

PROCEEDINGS

The HKBU 7th Computer Science Postgraduate Research Symposium

January 10, 2008

PG Day 2008



**Department of Computer Science
Hong Kong Baptist University**

The 7th HKBU-CSD Postgraduate Research Symposium (PG Day) Program

January 10 Thursday, 2008			
Time	Sessions		
08:40-09:10	On-site Registration		
09:10-09:20	Welcome: Prof. Jiming LIU, Head of Computer Science Department (LMC 514)		
09:20-10:50	Session A1: (Chair: Jian LI) (LMC 514) Data Mining <ul style="list-style-type: none"> • <i>Forum Classification Using Semi-supervised Gaussian Processes</i> Chi Wa CHENG • <i>The Local Kernel Regression Score for Feature Selection</i> Hong ZENG • <i>A Survey to Community Mining Algorithms in Complex Networks</i> Dan ZHANG 		
10:50-11:00	Tea Break		
11:00-12:00	Keynote Talk: Prof. Xindong Wu, University of Vermont, USA, and Hong Kong Polytechnic University, China (LMC 514) <ul style="list-style-type: none"> • <i>Top-10 Algorithms in Data Mining</i> 		
12:00-14:00	Noon Break		
14:00-15:30	<table style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 50%; vertical-align: top;"> Session A2: (Chair: Xiaofeng ZHANG) (LMC 512) Data Mining <ul style="list-style-type: none"> • <i>Latent Topics Detection with Dirichlet Process Mixture Model</i> Tianjie ZHAN • <i>A Concurrent G-Negotiation Mechanism for Grid Resource Co-allocation</i> Benyun SHI • <i>Privacy Policy Enforcement in Service-Oriented Data Analysis Workflows</i> Kai Kin CHAN </td> <td style="width: 50%; vertical-align: top;"> Session B: (Chair: Junyang ZHOU) (LMC 514) Information System <ul style="list-style-type: none"> • <i>Goal Oriented Requirements Engineering-Goal Definition</i> Di WU • <i>Automatic Semantic Annotation of Web Images</i> Chun Fan WONG • <i>Concept-based Multimedia Searching and Indexing</i> Wing Sze CHAN </td> </tr> </table>	Session A2: (Chair: Xiaofeng ZHANG) (LMC 512) Data Mining <ul style="list-style-type: none"> • <i>Latent Topics Detection with Dirichlet Process Mixture Model</i> Tianjie ZHAN • <i>A Concurrent G-Negotiation Mechanism for Grid Resource Co-allocation</i> Benyun SHI • <i>Privacy Policy Enforcement in Service-Oriented Data Analysis Workflows</i> Kai Kin CHAN 	Session B: (Chair: Junyang ZHOU) (LMC 514) Information System <ul style="list-style-type: none"> • <i>Goal Oriented Requirements Engineering-Goal Definition</i> Di WU • <i>Automatic Semantic Annotation of Web Images</i> Chun Fan WONG • <i>Concept-based Multimedia Searching and Indexing</i> Wing Sze CHAN
Session A2: (Chair: Xiaofeng ZHANG) (LMC 512) Data Mining <ul style="list-style-type: none"> • <i>Latent Topics Detection with Dirichlet Process Mixture Model</i> Tianjie ZHAN • <i>A Concurrent G-Negotiation Mechanism for Grid Resource Co-allocation</i> Benyun SHI • <i>Privacy Policy Enforcement in Service-Oriented Data Analysis Workflows</i> Kai Kin CHAN 	Session B: (Chair: Junyang ZHOU) (LMC 514) Information System <ul style="list-style-type: none"> • <i>Goal Oriented Requirements Engineering-Goal Definition</i> Di WU • <i>Automatic Semantic Annotation of Web Images</i> Chun Fan WONG • <i>Concept-based Multimedia Searching and Indexing</i> Wing Sze CHAN 		
15:30-15:40	Tea Break		
15:40-17:40	<table style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 50%; vertical-align: top;"> Session C: (Chair: Yicheng FENG) (LMC 512) Pattern Recognition <ul style="list-style-type: none"> • <i>Path Planning of Virtual Human by Reinforcement Learning</i> Yuesheng HE • <i>Tensor Locality Preserving Projections for Face Recognition</i> Limin CUI • <i>Moving Object Detection Based on Information Theoretic Spatio-Temporal Saliency</i> Chang LIU </td> <td style="width: 50%; vertical-align: top;"> Session D: (Chair: Zhili WU) (LMC 514) Database & Networking <ul style="list-style-type: none"> • <i>Processing Spatial Queries in Wireless Sensor Networks</i> Yu LI • <i>Data Management on Flash Storage</i> Saitung ON • <i>SepRep: A Novel Reputation Evaluation Model in Peer-to-Peer Networks</i> Xiaowei CHEN • <i>Performance Evaluation and Improvement of IEEE 802.11 Infrastructure Mode with Intra-Cell UDP Traffic</i> Yong YAN </td> </tr> </table>	Session C: (Chair: Yicheng FENG) (LMC 512) Pattern Recognition <ul style="list-style-type: none"> • <i>Path Planning of Virtual Human by Reinforcement Learning</i> Yuesheng HE • <i>Tensor Locality Preserving Projections for Face Recognition</i> Limin CUI • <i>Moving Object Detection Based on Information Theoretic Spatio-Temporal Saliency</i> Chang LIU 	Session D: (Chair: Zhili WU) (LMC 514) Database & Networking <ul style="list-style-type: none"> • <i>Processing Spatial Queries in Wireless Sensor Networks</i> Yu LI • <i>Data Management on Flash Storage</i> Saitung ON • <i>SepRep: A Novel Reputation Evaluation Model in Peer-to-Peer Networks</i> Xiaowei CHEN • <i>Performance Evaluation and Improvement of IEEE 802.11 Infrastructure Mode with Intra-Cell UDP Traffic</i> Yong YAN
Session C: (Chair: Yicheng FENG) (LMC 512) Pattern Recognition <ul style="list-style-type: none"> • <i>Path Planning of Virtual Human by Reinforcement Learning</i> Yuesheng HE • <i>Tensor Locality Preserving Projections for Face Recognition</i> Limin CUI • <i>Moving Object Detection Based on Information Theoretic Spatio-Temporal Saliency</i> Chang LIU 	Session D: (Chair: Zhili WU) (LMC 514) Database & Networking <ul style="list-style-type: none"> • <i>Processing Spatial Queries in Wireless Sensor Networks</i> Yu LI • <i>Data Management on Flash Storage</i> Saitung ON • <i>SepRep: A Novel Reputation Evaluation Model in Peer-to-Peer Networks</i> Xiaowei CHEN • <i>Performance Evaluation and Improvement of IEEE 802.11 Infrastructure Mode with Intra-Cell UDP Traffic</i> Yong YAN 		
19:00	Best Paper & Best Presentation Awards Announcement via Email		

TABLE OF CONTENTS

Session A1: Data Mining

<i>Online-Forum Topic Detection Based on User Participation</i> -----	1
Chi Wa CHENG	
<i>The Local Kernel Regression Score for Feature Selection</i> -----	6
Hong ZENG	
<i>A Survey to Community Mining Algorithms in Complex Networks</i> -----	14
Dan ZHANG	

Session A2: Data Mining

<i>Latent Topics Detection with Dirichlet Process Mixture Model</i> -----	24
Tianjie ZHAN	
<i>A Concurrent G-Negotiation Mechanism for Grid Resource Co-allocation</i> -----	30
Benyun SHI	
<i>Privacy Policy Enforcement in Service-Oriented Data Analysis Workflows</i> -----	34
Kai Kin CHAN	

Session B: Information System

<i>Goal Oriented Requirements Engineering-Goal Definition</i> -----	40
Di WU	
<i>Automatic Semantic Annotation of Web Images</i> -----	48
Chun Fan WONG	
<i>Concept-based Multimedia Searching and Indexing</i> -----	57
Wing Sze CHAN	

Session C: Pattern Recognition

<i>Path Planning of Virtual Human by Reinforcement Learning</i> -----	61
Yuesheng HE	
<i>Tensor Locality Preserving Projections for Face Recognition</i> -----	68
Limin CUI	
<i>Moving Object Detection Based on Information Theoretic Spatio-Temporal Saliency</i> -----	73
Chang LIU	

Session D: Database & Networking

<i>Processing Spatial Queries in Wireless Sensor Networks</i> -----	81
Yu LI	
<i>Data Management on Flash Storage</i> -----	89
Saitung ON	
<i>SepRep: A Novel Reputation Evaluation Model in Peer-to-Peer Network</i> -----	94
Xiaowei CHEN	
<i>Performance Evaluation and Improvement of IEEE 802.11 Infrastructure Mode with Intra-Cell UDP Traffic</i> -----	101
Yong YAN	

Forum Classification Using Semi-supervised Gaussian Processes

Victor Cheng

Abstract

The advent of Web2.0 enables the proliferation of online communities that information is mostly contributed and shared by enormous number of Internet users rather than web site owners or small groups of experts. Effective utilization of these resources can be achieved by classifying or clustering the contents. However, supervised classification requires labeling of patterns which is very expensive and time consuming. In this paper, we propose a semi-supervised Gaussian processes for forum discussion classification which employs the "null category noise model" (NCNM) to draw decision boundaries through low density pattern regions. Our experiments with forum content classification and handwritten digits classification show that high accuracy can be obtained with small number of labeled patterns.

1 Introduction

Recent years have seen a greatly increased attention to online communities where internet users play a dominant role in content contribution and sharing. From online-discussion to Wikipedia to weblogs and YouTube, they are featured with contributions of tremendous internet users all over the world rather than a limited set of professionals or authorities. Study of these users centric phenomena can be beneficial to a better understanding of the psychology and sociology of the communities which is significant to their healthy development. Kollock [5] studied the motivations of users in online forums and Bishop [1] investigated the methods to encourage participation. Social network analysis [2] had also been used to analyze various online communities and studied the relationship and roles of network users. Nolker and Zhou [8] applied social network theory to newsgroups to identify leaders, motivators and chatters.

Recently, classification of forum discussions has become an increasingly important area of research in information retrieval. While much success has been obtained in classifying news broadcast, e.g. Google News, progress in online forum is hindered by the imprecise, terse and causal communication styles. Previous research work on analyzing con-

tents in online forum often treated an online article as individual text document without considering posters' posting characteristics. Cheng [3] used participation frequencies as attributes and found that they were effective in topic detection of web forums. In this paper, we show that user participation can also facilitate supervised learning and even semi-supervised learning. Semi-supervised paradigm is important in understanding forum discussion because getting millions of discussions is now not difficult but labeling them properly is very costly and time consuming and sometimes it is regarded as impossible.

There are many formulations available for semi-supervised classification. Traditional approaches such as labels propagation algorithm [11] and transductive support vector machine (T-SVM) [10] are very popular and simple to use. We propose Lawrence's [7] semi-supervised Gaussian learning for forum discussion classification. It is because not only the unlabelled data can be exploited but also further information embedding can be done easily through Bayes' formula, through the likelihood function (or called the noise model). For example, if f is the latent variable of a Gaussian processes [9] and I is additional information we want to embed, which is modeled by a probability function $p(I)$, the posterior distribution of f conditioned on I is given by

$$p(f|I) = \frac{p(I|f)p(f)}{p(I)}. \quad (1)$$

Hence, through this pulp-and-play formulation, gaussian processes can embed additional information systematically. On the other hand, it is not trivial to embed new information to support Vector Machine (SVM). Information can be embedded only by changing the cost functions and the corresponding constraints of the quadratic optimization problem during formulating the SVM and there are no systematic ways developed for doing so.

In semi-supervised learning, the additional information we obtained is a vast amount of unlabelled patterns, \mathbf{x} . If we fail to make an assumption about the underlying distribution, $p(\mathbf{x})$, the unlabelled patterns will be useless and has no effects to the classification results. A weak and reasonable assumption made in this paper is that the probability distribution of patterns is relatively sparse near the bound-

ary of classes. This assumption is in fact made use by the T-SVM which seeks the maximum margin by allocating labels to the unlabelled patterns. In probability framework, this low pattern density near decision boundary assumption can be modeled by null category noise model [7]. Our Experiments on a local web forum and USPS handwritten digits show that the performance is comparable to T-SVM and sometimes it works a little bit better.

The paper is organized as follow. In section 2, we describe the semi-supervised Gaussian processes. Experimental results are given in Section 3 and Section 4 is the Conclusion.

2 Semi-supervised Gaussian Processes

In the standard setting of inductive learning, a pattern is described by a collection of input attributes, denoted by a column vector $\mathbf{x} \in \chi \subset \mathbf{R}^d$. The key idea of Gaussian process models is to introduce a random variable, called latent variable, $f_{\mathbf{x}}$ for each point \mathbf{x} in χ . These random variables $\{f_{\mathbf{x}}\}_{\mathbf{x} \in \chi}$ are treated as outputs of a zero-mean Gaussian process, or more exactly a zero-mean multivariate Gaussian distribution. The covariance between two random variable $f_{\mathbf{x}}$ and $f_{\mathbf{z}}$ is defined by a Mercer kernel function $K(\mathbf{x}, \mathbf{z})$. Thus, the prior distribution over $\mathbf{f} = [f_{\mathbf{x}_1} \dots f_{\mathbf{x}_n}]$ associated with a collection of n points, $\mathbf{x}_1 \dots \mathbf{x}_n$, can be written as

$$p(\mathbf{f}) = \frac{1}{(2\pi)^{n/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2} \mathbf{f}^T \Sigma^{-1} \mathbf{f}\right) \quad (2)$$

where Σ is the $n \times n$ covariance matrix whose ij -th element is $K(\mathbf{x}_i, \mathbf{x}_j)$. Under the context of binary classification, the goal is to find a mapping between the inputs $\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_n]$ and labels $Y = [y_1 \dots y_n]$, $y_i \in \{-1, 1\}$, which is given by

$$p(Y|\mathbf{X}) = \int p(Y|\mathbf{f})p(\mathbf{f}|\mathbf{X})d\mathbf{f}. \quad (3)$$

The term $p(y_i|f_i)$ in $p(Y|\mathbf{f})$ is called the likelihood or the noise model. In semi-supervised learning formulation of Gaussian processes, we employ "null category noise model" (NCNM) in which y_i in $p(y_i|f_i)$ can take one value in $\{-1, 0, 1\}$ instead of $\{-1, 1\}$. The noise process model is described by

$$p(y_i|f_i) = \begin{cases} \phi(-(f_i + w/2)) & \text{for } y_i = -1 \\ \phi(f_i + w/2) - \phi(f_i - w/2) & \text{for } y_i = 0 \\ \phi(f_i - w/2) & \text{for } y_i = 1 \end{cases} \quad (4)$$

where $\phi(x) = \int_{-\infty}^x N(z|0, 1)dz$ is the cumulative Gaussian distribution function and w is a parameter controlling

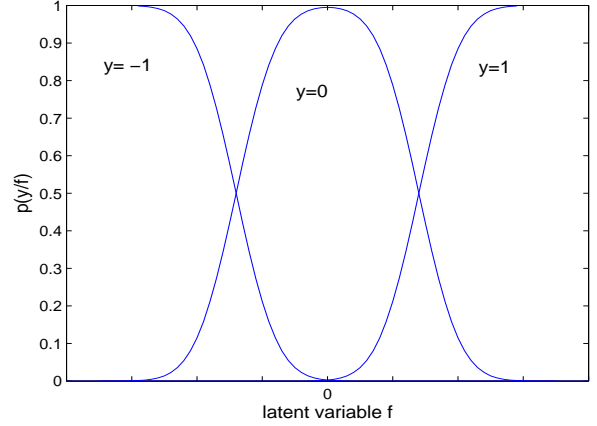


Figure 1. The 3 categories noise process model.

the width of the category $y_i = 0$. The interpretation of this model is intuitive. If f_i is not at the vicinity of zero, e.g. positive/negative enough, $p(y = 1|f_i)/p(y = -1|f_i)$ approaches 1.0/-1.0, otherwise $p(y = 0|f_i)$ will be dominant. In fact, the region corresponds to $y = 0$ is the region near the decision boundary. To use this model in an unlabeled setting, an additional random variable z_i is introduced which has the value one if the pattern \mathbf{x}_i is unlabeled and zero otherwise. As discussed in Section 1, the assumption that the decision boundary are drawn through low density regions can thus be implemented by imposing the constraint

$$p(z_i = 1|y_i = 0) = 0. \quad (5)$$

Thus, an unlabeled pattern should not come from category $y = 0$. This noise model prefers choosing a decision boundary which has category $y = 0$ having less unlabeled patterns in the neighborhood, and it should through regions with low pattern density. The resultant noise model is obtained by combining (4) and (5). To ease the computation, the cumulative Gaussian distribution in (4) is replaced by the step function $H(\cdot)$. If the class proportion, γ_+, γ_- , of unlabeled patterns is known,

$$p(z_i = 1|y_i = 1) = \gamma_+ \quad (6)$$

$$p(z_i = 1|y_i = -1) = \gamma_-, \quad (7)$$

the resultant noise model can be simplified into

$$p(y_i|f_i) = \begin{cases} H(-(f_i + 1/2)) & \text{for } y_i = -1, z_i = 0 \\ \gamma_- H(-(f_i + 1/2)) + \gamma_+ H(f_i - 1/2) & \text{for } z_i = 1 \\ H(f_i - 1/2) & \text{for } y_i = 1, z_i = 0 \end{cases}, \quad (8)$$

where we take $w = 1/2$.

2.1 Posterior Computation and Prediction

Combining the Gaussian process prior (\mathbf{f}) with the noise model ($p(y_i|f_i)$), the posterior distribution is given as follows,

$$p(\mathbf{f}|Y) = \frac{1}{p(Y)} p(\mathbf{f}) \prod_{i=1}^n p(y_i|f_i). \quad (9)$$

However, direct computation is intractable because the noise model is no longer Gaussian. To preserve computational tractability, the posterior distribution is approximated by a joint Gaussian centered at the true mean. A family of inference techniques can be applied for the Gaussian approximation, e.g. Laplace approximation or expectation propagation. In this paper, we employ the "assumed density filtering" (ADF) method [6] because of its simplicity. To apply ADF, the prior is first approximated by a Gaussian distribution $q_0(\mathbf{f}|\mu_0, \Sigma_0)$. Since the prior $p(\mathbf{f})$ is Gaussian, the approximation is exact. After incorporating the first factor of the noise model, the exact posterior $\hat{p}_1(\mathbf{f}|y_1)$ is given by

$$\hat{p}_1(\mathbf{f}|y_1) = \frac{1}{Z_1} q_0(\mathbf{f}) p(y_1|f_1), \quad (10)$$

where the normalization constant is

$$Z_1 = \int q_0(\mathbf{f}) p(y_1|f_1) d\mathbf{f}. \quad (11)$$

This posterior is then approximated by $q_1(\mathbf{f}|\mu_1, \Sigma_1)$ by minimizing the KL divergence between it and $\hat{p}_1(\mathbf{f}|y_1)$,

$$KL(\hat{p}_1(\mathbf{f}|y_1)||q_1(\mathbf{f})) = \int \hat{p}_1(\mathbf{f}|y_1) \ln \frac{\hat{p}_1(\mathbf{f}|y_1)}{q_1(\mathbf{f})} d\mathbf{f}. \quad (12)$$

After getting $q_1(\mathbf{f})$, the next factor of the noise model is included to form $\hat{p}_2(\mathbf{f}|y_1, y_2)$ and the approximation is repeated and we get $q_2(\mathbf{f})$. This is continued until all training patterns are included.

During the inclusion of the noise factor $p(y_i|f_i)$ of the training pattern \mathbf{x}_i , the minimization of the KL divergence can be computed by moment matching. As

$$\mu_i = E_{\hat{p}_i(\mathbf{f})}(\mathbf{f}) \quad (13)$$

$$\Sigma_i = E_{\hat{p}_i(\mathbf{f})}(\mathbf{f}\mathbf{f}^T) - \mu_i\mu_i^T, \quad (14)$$

the computation depends on the tractability of the normalization constant in (11). To save the space, we just quote the final updating equations for the posterior mean and covariance here. Interested readers can refer [6] for detailed description.

$$\mu_i = \mu_{i-1} + \Sigma_{i-1}\mathbf{g}_i, \quad (15)$$

where \mathbf{g}_i is given by

$$\mathbf{g}_i = \frac{\partial \log Z_i}{\partial \mu_i}. \quad (16)$$

$$\Sigma_i = \Sigma_{i-1} - \Sigma_{i-1}(\mathbf{g}_i\mathbf{g}_i^T - 2\Gamma_i)\Sigma_{i-1} \quad (17)$$

where

$$\Gamma_i = \frac{\partial Z_i}{\partial \Sigma_{i-1}}. \quad (18)$$

In general, the ADF approximation is not limited to the noise model described here and can be applied for any noise models for which the computation of (11) is tractable.

Label prediction of a testing point \mathbf{x}_t can be done via the marginal distribution $p(y_t|\mathbf{x}_t)$. An issue is raised since NCM has a non-zero probability state $y = 0$, which should not exist in reality. This is where the role of z becomes essential again. We set $z_t = 1$ so in reality the probability that this data point is from the positive class is given by

$$p(y_t|\mathbf{x}_t, z_t) \propto p(z_t|y_t)p(y_t|\mathbf{x}_t). \quad (19)$$

The constraint that $p(z_t = 1|y_t = 0) = 0$ causes this point coming from $y = 0$ has probability zero and hence the issue can be resolved.

3 Experiments

In this section, we present the results of applying the semi-supervised Gaussian processes to classify discussions of a local Audio-Visual web forum and handwritten digits from the USPS dataset [4]. In each experiment, we compare the performance of SVM, T-SVM, Gaussian processes, semi-supervised Gaussian processes under different proportions of labeled and unlabeled patterns.

3.1 Classification of Discussions of a Local Forum

The web forum which used in the experiment provides a discussion cyberspace for people interested in Audio-visual affairs, in particular the high-end or high fidelity (Hi-Fi) equipment. To avoid any advertising effects, the alias **AVForum** is used. In this forum, three distinct discussion boards are available to public users with assigned alias **AvBoard**, **ChatBoard**, and **2ndHandBoard**. In **AvBoard** users are welcome to share their idea on Audio-visual affairs, **ChatBoard** provides a space for unbounded casual chats (except illegal affairs), and **2ndHandBoard** is a platform for people posting advertisements for buying or selling 2nd hand products. There are totally 7728 distinct user names and they form the attributes of all discussions. Discussions with few users are removed and results in 1003 discussions in **ChatBoard**, 1069 discussions in **2ndHandBoard**, and 1039 discussions in **AvBoard**. Classifications are done by first randomly sampling 400 discussions from any two discussion boards (i.e. 200 discussions from each board) and the rest discussions of them are mixed playing the role of testing patterns. Among the 400 sampled discussions, a proportion

of them are labeled and all other are treated as unlabeled training patterns. The process is repeated for five times and the means of the results are recorded. Figure 2 and 3 show the classification results in AUC, the area under the ROC curve. AUC is used in this experiment rather than accuracy because it is found that the output of the SVM and T-SVM is biased and using zero as threshold value for class discrimination is not fair.

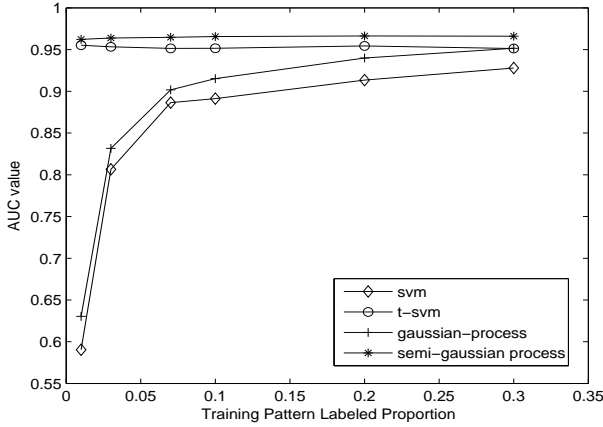


Figure 2. ChatBoard versus 2ndHandBoard classification results.

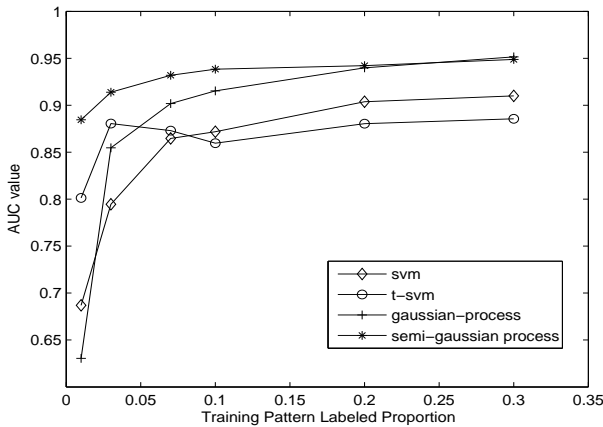


Figure 3. ChatBoard versus AVBoard classification results.

From the figures, it is clear that the benefits from unlabeled patterns become significant when the labeled proportion of the training patterns below 0.1 (below 40 discussions). The performance of both Gaussian processes seem to be a little better than SVM and T-SVM,

respectively. It is also shown that the semi-supervised Gaussian processes and T-SVM is less sensitive to the proportion of labeled data and produce high AUC even there are only a small number of labeled patterns. This feature is very important because it is usually a huge amount of discussions encountered in analyzing forums and semi-supervised learning seems to be one of the possible ways to deal with them.

3.2 Classification of USPS Dataset

The second experiment is the classification of handwritten digits of the well known USPS dataset. It originally consists of a training set with 7291 images and a test set with 2007 images. We mix and reshuffle all the images and divided them into equal size of training set and testing set with each of them contains 4649 cases. After the process, both the training set and testing set contain about 300-500 images for each digit between 0-9. Each digit image contains a raster scan of the 16×16 grey level pixel intensities which are scaled such that the range is between -1 and 1. The classification results for the digits 3 versus 5 and 7 versus 9 are shown in the figure 4 and 5 respectively. In the experiment, the quality of classification is measured in error rate as the output bias problem described in the previous subsection does not exist.

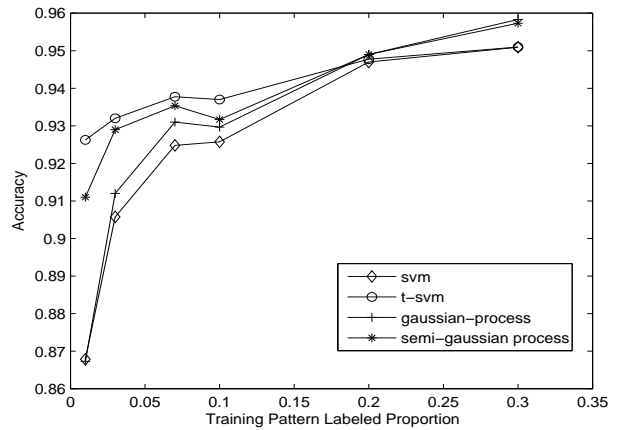


Figure 4. Digit 3 versus 5 classification results.

From the figures, it is still seen that the semi-supervised Gaussian processes and T-SVM work better than their respective standard versions especially when the labeling proportion is small. These results also show that semi-supervised Gaussian processes performs similarly as T-SVM. As we have described in Section 1, semi-supervised Gaussian processes is preferred because additional informa-

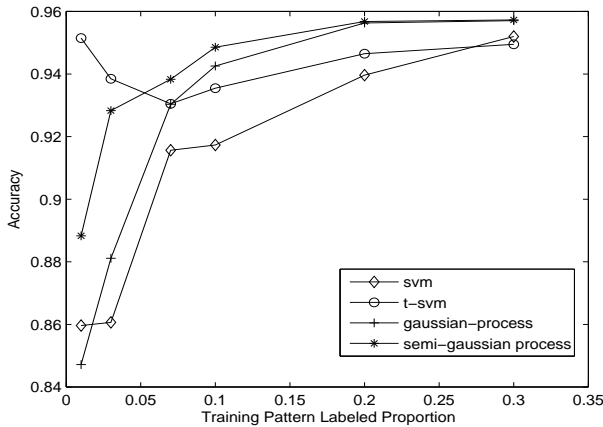


Figure 5. Digit 7 verses 9 classification results.

tion can be embedded into the model easily and in a systematic manner.

4 Conclusion

Nowadays, internet users are playing more and more important roles in content contribution and social networks. Classifying the contents they created is helpful for analysis and healthy development. This task is not easy. For example, it is not difficult to grep a huge amount of discussions but labeling them is another story. In this paper, we propose using semi-supervised Gaussian processes to classify discussions. Semi-supervised Gaussian processes is chosen rather than SVM or T-SVM because its performance is comparable to SVM and sometime a little bit better. The most important feature of it is that additional information can be embedded into the model in an easy and systematic way. It is also found that using user participation as attributes is also a good choice other than words. From the experimental results, high accuracy can be achieved with just participation information. The future direction will be the investigation of combining words and participation for tasks such as clustering and classification. This also benefits the study and revealing the relations between people and their interests.

Acknowledgements

This work is supported by the Research Grant Council Central Allocation Fund of the Hong Kong SAR Government under the grant number: HKBU 1/05C.

References

- [1] J. Bishop, "Increasing participation in online communities: A framework for human-computer interaction," *Comput. Hum. Behav.*, 23(4):pages 1881-1893, 2007.
- [2] P.J. Carrington, J. Scott, and S. Wasserman, *Models and Methods in Social Network Analysis*, Cambridge University Press, 2005.
- [3] V. Cheng, Z. Wu, and C.H. Li., "Online-Forum Topic Detection Based on User Participation," in *Proc.IEEE ICEBE 2007 Student Workshop*, 18-23, 2007.
- [4] J. J. Hull, "A Database for Handwritten Text Recognition Research," *IEEE PAMI*, 16(5) 550-554, 1994.
- [5] P. Kollock. "The economies of online cooperation: Gifts and public goods in cyberspace," *Communities in Cyberspace*, Routledge, London, 1999.
- [6] N.D. Lawrence, J.C. Platt, and M.I. Jordan, "Extensions of the informative vector machine," in *Deterministic and Statistical Methods in Machine Learning*, Springer-Verlag, Berlin, 2005.
- [7] N.D. Lawrence and M.I. Jordan, "Semi-supervised learning via Gaussian processes," in *Advances in Neural Information Processing Systems*, MIT Press, Cambridge, MA., 2005.
- [8] R.D. Nolker and L. Zhou, "Social computing and weighting to identify member roles in online communities," in *WI'05: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 87-93. IEEE Computer Society, 2005.
- [9] C.E. Rasmussen and C.K.I. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.
- [10] V.N. Vapnik, *Statistical Learning Theory*. John Wiley and Sons, New York, 1998.
- [11] X. Zhu, *Semi-Supervised Learning with Graphs*. Doctoral thesis, CMU-LTI-05-192, May 2005

The Local Kernel Regression Score for Feature Selection

Hong Zeng

Abstract

In many unsupervised and supervised learning tasks, the input is often represented by a very large number of features, but many of them may be irrelevant to the tasks. Feature selection is an effective tool to identify the relevant features and reduce the dimensionality. This paper presents a Local Kernel Regression (LKR) scoring approach, it selects the features by ranking their capabilities to hold the pre-extracted local similarities among data in a small patch, which are assumed to be crucial for the discrimination. The evaluation is done by investigating how well the feature value of each point could be estimated by its neighbors and their feature values, within a local kernel regression framework. Experimental results show the effectiveness of the proposed method.

1 Introduction

In many domains of machine learning applications, *e.g.*, the computer vision, the text classification and the more recent gene expression array analysis, *etc.*, the input raw data is often very high dimensional, and may have many irrelevant features, the problem of focusing on the most informative features of the data has become increasingly important. Feature selection is an effective tool to identify the relevant features and reduce the dimensionality of data for the inference tasks [4, 7].

Three categories of the feature selection methods are often distinguished [1]: *filter*, *wrapper* and *embedded* approaches. The *filter* methods usually apply the feature selection almost independently of the succeeding inference schemes. The *wrapper* and *embedded* approaches utilize the intermediate outputs of the employed learning algorithm to evaluate the quality of the feature subset. Though good predictors generally could be obtained by the *wrapper* and *embedded* methods, they are often criticized for highly computational expense, due to repeatedly wrapping the subset selection around the learning algorithm (“*wrapper*”) or requiring a laborious iterative optimization process (“*embedded*”). Hence we are particularly interested in developing a *filter* method which may be more appropriate for

dealing with the high-dimensional data.

In the literature, with the class labels involved in the feature selection process, the supervised *filter* approaches have been largely studied, *e.g.*, χ^2 -test, Fisher score, Information Gain, *etc.* By contrast, the unsupervised counterparts, which are without the participation of labels in the selection, have received less attention. However, for clustering and other scenarios where the labeled data may be unable or costly to obtain, it is also very crucial to distinguish the representative features that could induce better discrimination for the partitioning. The main difficulty of developing the unsupervised feature selection lies in the absence of class labels which could guide the process. Until recently, several unsupervised *filter* methods, trying to select the relevant features based on the intrinsic properties of the data, have been proposed. Wolf *et al.* [13] proposed the Q - α algorithm, which builds on the spectral properties of the graph laplacian of data on the candidate feature subset, and iteratively calculates the soft cluster indicator matrix and the feature weights. Although an interesting property of sparsity in feature weights naturally emerges, the iterative optimization may not scale well in case of thousands of features. A more computation-saving method, Laplacian score, is proposed by He *et al.* [5]. It also takes advantage of the graph laplacian, but selects the features by ranking their capabilities of preserving the locality in the graph. Zhao *et al.* [17] developed a more general spectral feature selection framework, which includes the Laplacian score as a special case.

In this paper, we present a novel feature selection method, called *Local Kernel Regression* (LKR) score, which could be performed in both unsupervised and supervised manner to select the representative features from the high-dimensional data. The proposed approach is fundamentally based on the kernel regression and the nearest neighborhood graph. Specifically, for the points from a small patch in the graph, it is reasonable to assume that the feature value of each point can be estimated based on its neighbors and their feature values. Thus the estimation error could reflect the ability of the feature to preserve the local relationship among the data points, which is assumed to be crucial for the discrimination. We adopt a local kernel regression model to conduct the estimation and will show that the state-of-the-art Laplacian score [5] feature selection

method can be also interpreted from a local kernel regression perspective.

The remainder of the paper is organized as follows. In Section 2, we introduce the proposed Local Kernel Regression (LKR) score algorithm for feature selection. Some connections with related methods are given in Section 3. The extensive experimental results are presented in Section 4. Finally, we conclude the paper in Section 5.

2 The local kernel regression score for feature selection

The notions used in this study are introduced first. $\mathcal{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$, $\mathbf{x}_i \in \mathbb{R}^d$ represents a data set of N samples. $\mathbf{y} = [y_1, \dots, y_N]^T$ denotes the vector of labels. \mathcal{N}_i denotes the neighboring set of points of \mathbf{x}_i , $1 \leq i \leq N$. $n_i = |\mathcal{N}_i|$ is the number of neighboring points of \mathbf{x}_i . We use F_1, \dots, F_d to denote the d features, and let $\Phi_{LKR}(F_l)$ denote the LKR score of the l -th feature vector \mathbf{f}_l , $\mathbf{f}_l = [x_1^{(l)}, \dots, x_N^{(l)}]^T$, the i -th element in \mathbf{f}_l is denoted as $\mathbf{f}_l^{(i)}$.

Before introducing the main algorithm, we briefly review the kernel regression algorithm adopted in the paper, *i.e.*, the kernel ridge regression, and then the nearest neighborhood graph.

2.1 Kernel ridge regression

Kernel ridge regression is a simple yet very effective tool for building non-linear regression model [11]. Given the training data $\{(\mathbf{x}_i, t_i)\}_{i=1}^N$, where $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ is the input data, and $t_i \in \mathbb{t} \subset \mathbb{R}$ is the real valued target, it seeks a function that could map \mathcal{X} to \mathbb{R} under the least-squares framework. The predictive model of kernel ridge regression can be expressed as:

$$g(\mathbf{x}) = \sum_{i=1}^N \alpha_i \mathcal{K}(\mathbf{x}, \mathbf{x}_i) \quad (1)$$

where α_i 's are the estimation coefficients, and $\mathcal{K} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is a kernel function. The α_i 's are solved by minimizing the following objective function which consists of the fitness item and the Tikhonov regularization item [12]:

$$\|\mathbf{K}\alpha - \mathbf{t}\|^2 + \lambda \alpha^T \mathbf{K}\alpha \quad (2)$$

where $\alpha = [\alpha_1, \dots, \alpha_N]^T \in \mathbb{R}^N$, λ is a small positive regularization parameter, $\mathbf{t} = [t_1, \dots, t_N]^T$ denotes the vector of real valued targets, and $\mathbf{K} = [k_{ij}]_{N \times N}$, ($k_{ij} = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$) is the kernel matrix over all the training samples \mathcal{X} . By setting the derivatives of the objective function with respect to α equal to the zero vector, we could obtain the solution of (2) given by:

$$\alpha = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{t} \quad (3)$$

where \mathbf{I} is a $N \times N$ unit matrix. Finally the kernel ridge regression model can be expressed as:

$$g(\mathbf{x}) = \mathbf{k}_x^T \alpha = \mathbf{k}_x^T (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{t} \quad (4)$$

where $\mathbf{k}_x = [\mathcal{K}(\mathbf{x}, \mathbf{x}_1), \dots, \mathcal{K}(\mathbf{x}, \mathbf{x}_N)]^T \in \mathbb{R}^N$.

2.2 The nearest neighborhood graph

Given the data set \mathcal{X} , let $\mathbf{G}(\mathbf{V}, \mathbf{E})$ be the undirected graph constructed from \mathcal{X} , with \mathbf{V} being its node set and \mathbf{E} being its edge set. The i -th node v_i of \mathbf{G} corresponds to $\mathbf{x}_i \in \mathcal{X}$. The \mathbf{G} is constructed as follows: if v_i is in the *neighborhood* of v_j , or v_j is in the *neighborhood* of v_i ($i \neq j$), an edge is put between node i and j . For the unsupervised case, the *neighborhood* of v_i can be defined as its k nearest neighbors (excluding v_i itself) of a data according to certain distance metric, such as the Euclidean distance used in this paper. While for the supervised case, the *neighborhood* could be defined as the nodes which share the same class labels. Let \mathbf{W} be the symmetric $N \times N$ weight matrix, with w_{ij} being the weight of the edge joining vertices i and j . For the unsupervised case, the weight w_{ij} can be calculated by:

$$w_{ij} = \begin{cases} \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j), & \text{if } \mathbf{x}_i \in \mathcal{N}_j \text{ or } \mathbf{x}_j \in \mathcal{N}_i; \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

where the $\mathcal{K}(\cdot, \cdot)$ is the kernel function mentioned above. For the supervised case, the following form can be adopted:

$$w_{ij} = \begin{cases} \frac{1}{N_l}, & \text{if } y_i = y_j = l; \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

where N_l denotes the number of data points in class l . The degree matrix \mathbf{D} of the graph \mathbf{G} is defined by: $D_{ij} = \sum_{m=1}^N w_{im}$ if $i = j$, and 0 otherwise. According to the spectral graph theory [2], the density around \mathbf{x}_i could be approximated by D_{ii} , and the more points are close to \mathbf{x}_i , the larger is D_{ii} .

2.3 The local kernel regression score

Since the within-cluster and within-class similarities are very useful for the discrimination, it is reasonable to select the features that could keep such similarities or configurations within a small patch on the graph. Such a criterion is implemented by examining how well the feature value of each point can be estimated based on its neighbors and their feature values. After picking out the relevant features, the intra-similarity within the same cluster or class could become more distinct, which will undoubtedly help the partitioning task.

In this work, we employ a local kernel ridge regression to implement the estimation. Given the input data \mathbf{x}_i and

$\{(\mathbf{x}_j, \mathbf{f}_l^{(j)})\}_{\mathbf{x}_j \in \mathcal{N}_i}$, we want to train a kernel ridge regression model locally to approximate the l -th feature value of \mathbf{x}_i (*i.e.* $\mathbf{f}_l^{(i)}$), with the training data $\{(\mathbf{x}_j, \mathbf{f}_l^{(j)})\}_{\mathbf{x}_j \in \mathcal{N}_i}$, where $\mathbf{f}_l^{(j)}$ is used as the real-valued target of \mathbf{x}_j for learning this kernel machine. Based on equation (4), we use the following equation to denote the local kernel ridge regression model at \mathbf{x}_i :

$$g_{\mathcal{N}_i}(\mathbf{x}_i) = \mathbf{k}_{\mathcal{N}_i}^T (\mathbf{K}_{\mathcal{N}_i} + \lambda \mathbf{I})^{-1} \mathbf{f}_l^{(\mathcal{N}_i)} \quad (7)$$

where $g_{\mathcal{N}_i}(\cdot)$ denotes the regression model learned with the training data $\{(\mathbf{x}_j, \mathbf{f}_l^{(j)})\}_{\mathbf{x}_j \in \mathcal{N}_i}$, $\mathbf{k}_{\mathcal{N}_i} \in \mathbb{R}^{n_i}$ denotes the vector $[\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)]^T$ for $\mathbf{x}_j \in \mathcal{N}_i$, $\mathbf{f}_l^{(\mathcal{N}_i)} \in \mathbb{R}^{n_i}$ denotes the vector $[\mathbf{f}_l^{(j)}]^T$ for $\mathbf{x}_j \in \mathcal{N}_i$, $\mathbf{K}_{\mathcal{N}_i} \in \mathbb{R}^{n_i \times n_i}$ is the kernel matrix over $\mathbf{x}_j \in \mathcal{N}_i$, *i.e.*, $\mathbf{K}_{\mathcal{N}_i} = [\mathcal{K}(\mathbf{x}_p, \mathbf{x}_q)]$, for $\mathbf{x}_p, \mathbf{x}_q \in \mathcal{N}_i$, and \mathbf{I} is a $n_i \times n_i$ unit matrix. We let

$$\beta_{\mathcal{N}_i}^T = \mathbf{k}_{\mathcal{N}_i}^T (\mathbf{K}_{\mathcal{N}_i} + \lambda \mathbf{I})^{-1} \quad (8)$$

where $\beta_{\mathcal{N}_i} \in \mathbb{R}^{n_i}$, then the equation (7) can be rewritten in a linear form:

$$g_{\mathcal{N}_i}(\mathbf{x}_i) = \beta_{\mathcal{N}_i}^T \mathbf{f}_l^{(\mathcal{N}_i)} \quad (9)$$

We now introduce a new vector $\beta_i \in \mathbb{R}^N$, and the j -th ($1 \leq j \leq N$) element β_{ij} is calculated as follows: if $\mathbf{x}_j \in \mathcal{N}_i$, then β_{ij} equals to the corresponding element of $\beta_{\mathcal{N}_i}$ in (8), otherwise, it equals to 0. Note that $\mathbf{f}_l^{(\mathcal{N}_i)}$ is a sub-vector of \mathbf{f}_l , we rewrite (9) in a full vector form:

$$g(\mathbf{x}_i) = \beta_i^T \mathbf{f}_l = \sum_{j=1}^N \beta_{ij} \mathbf{f}_l^{(j)} \quad (10)$$

Therefore, the squared local estimation error for the l -th feature at \mathbf{x}_i is computed as:

$$E_{local}(\mathbf{f}_l^{(i)}) = (\mathbf{f}_l^{(i)} - g(\mathbf{x}_i))^2 = (\mathbf{f}_l^{(i)} - \sum_{j=1}^N \beta_{ij} \mathbf{f}_l^{(j)})^2 \quad (11)$$

Since the data density often varies over the whole data set. Some points may reside in a dense region, while others may lie in a sparse one, the importance of each point may not be the same. Therefore we compute the overall estimation error over the data manifold as a data density weighted sum:

$$\begin{aligned} E_{local}(\mathbf{f}_l) &\propto \sum_{i=1}^N (\mathbf{f}_l^{(i)} - \sum_{j=1}^N \beta_{ij} \mathbf{f}_l^{(j)})^2 D_{ii} \\ &= \sum_{i=1}^N [\sqrt{D_{ii}} \mathbf{f}_l^{(i)} - \sum_{j=1}^N (\beta_{ij} \sqrt{\frac{D_{ii}}{D_{jj}}}) \sqrt{D_{jj}} \mathbf{f}_l^{(j)}]^2 \\ &= \mathbf{f}_l^T \mathbf{D}^{\frac{1}{2}} (\mathbf{I} - \mathbf{B})^T (\mathbf{I} - \mathbf{B}) \mathbf{D}^{\frac{1}{2}} \mathbf{f}_l \end{aligned} \quad (12)$$

where $\mathbf{B} = [\beta_{ij} \sqrt{\frac{D_{ii}}{D_{jj}}}]_{N \times N}$ obviously will be a sparse matrix. Apparently, we should prefer the feature which could

achieve a small value for (12). However, one could easily find that the vector with all zero elements will get the smallest value (*i.e.*, 0), and this is an obvious trivial candidate; besides, the feature vector, whose elements are non-zero constants, also does not carry much information. In other words, a wide scattered band of the feature should also be required in order to have enough representative power [5]. It can be detected by investigating whether the feature has a large variance along the data manifold, and the density weighted variance can be estimated by:

$$\begin{aligned} Var_{\mathcal{M}}(\mathbf{f}_l) &\propto \sum_{i=1}^N (\mathbf{f}_l^{(i)} - \mu_{\mathcal{M}}(\mathbf{f}_l))^2 D_{ii} \\ &= (\mathbf{f}_l - \mu_{\mathcal{M}}(\mathbf{f}_l) \mathbf{e})^T \mathbf{D} (\mathbf{f}_l - \mu_{\mathcal{M}}(\mathbf{f}_l) \mathbf{e}) \\ &= \hat{\mathbf{f}}_l^T \mathbf{D} \hat{\mathbf{f}}_l. \end{aligned} \quad (13)$$

where $\mathbf{e} = [1, \dots, 1]^T \in \mathbb{R}^N$, $\hat{\mathbf{f}}_l = \mathbf{f}_l - \mu_{\mathcal{M}}(\mathbf{f}_l) \mathbf{e}$, $\mu_{\mathcal{M}}(\mathbf{f}_l)$ is the weighted mean of the l -th feature calculated by:

$$\begin{aligned} \mu_{\mathcal{M}}(\mathbf{f}_l) &= \sum_{i=1}^N \frac{D_{ii}}{\sum_{j=1}^N D_{jj}} \mathbf{f}_l^{(i)} \\ &= \frac{\sum_{i=1}^N D_{ii} \mathbf{f}_l^{(i)}}{\sum_{j=1}^N D_{jj}} \\ &= \frac{\mathbf{f}_l^T \mathbf{D} \mathbf{e}}{\mathbf{e}^T \mathbf{D} \mathbf{e}}, \end{aligned} \quad (14)$$

Eventually, we formulate the local kernel regression score as an integration of the two requirements (*i.e.* (12) and (13)) for a good feature:

$$\begin{aligned} \Phi_{LKR}(F_l) &= \frac{E_{local}(\mathbf{f}_l)}{Var_{\mathcal{M}}(\mathbf{f}_l)} \\ &= \frac{\mathbf{f}_l^T \mathbf{D}^{\frac{1}{2}} (\mathbf{I} - \mathbf{B})^T (\mathbf{I} - \mathbf{B}) \mathbf{D}^{\frac{1}{2}} \mathbf{f}_l}{\hat{\mathbf{f}}_l^T \mathbf{D} \hat{\mathbf{f}}_l}. \end{aligned} \quad (15)$$

We will seek the feature that the within-neighborhood estimation error is small, whereas its variance cross all data points is large. In practice, we rank the features¹ in ascending order of $\Phi_{LKR}(\cdot)$, and choose the top features appearing in the rank list. The four main steps of the proposed local kernel regression scoring algorithm are summarized in Algorithm 1.

3 Connections to related approaches

3.1 Connection to Laplacian score

The recently proposed Laplacian score [5] is an effective feature selection method, whose starting point is to seek

¹If \mathbf{f}_l is a constant vector, *e.g.*, $\mathbf{0}$ or \mathbf{e} , it could be obtained that the $Var_{\mathcal{M}}(\mathbf{f}_l) = 0$, therefore this trivial candidate can be easily excluded from the selection.

input : \mathcal{X} (and \mathbf{y}), the number of nearest neighbors

output: the ranked feature list

- 1 Construct the kernel matrix \mathbf{K} over \mathcal{X} ;
- 2 Construct the nearest neighborhood graph \mathbf{G} and the weight matrix \mathbf{W} (using (5) for unsupervised feature selection or (6) for supervised feature selection);
- 3 Learn the local kernel regression model with \mathbf{K} , \mathbf{G} by (8) and (10);
- 4 Compute LKR score for each feature by (15), then rank features in ascending order of $\Phi_{LKR}(F_l)$ ($l = 1, \dots, d$);

Algorithm 1: The Local Kernel Regression Score for Feature Selection.

the feature that could preserve the locality. Specifically, it prefers the features that could minimize the following objective function [5]:

$$\sum_{ij} (\mathbf{f}_l^{(i)} - \mathbf{f}_l^{(j)})^2 w_{ij} \quad (16)$$

where the w_{ij} is the similarity between \mathbf{x}_i and \mathbf{x}_j (the RBF kernel function is adopted for w_{ij} in [5]). It could be noted that a bigger $(\mathbf{f}_l^{(i)} - \mathbf{f}_l^{(j)})^2$ would produce a heavy cost if w_{ij} is large, thus indicating F_l is an undesirable feature. By setting the derivative of the equation (16) with respect to $\mathbf{f}_l^{(i)}$ to 0, we could find that minimizing $\sum_{ij} (\mathbf{f}_l^{(i)} - \mathbf{f}_l^{(j)})^2 w_{ij}$ requires the elements of \mathbf{f}_l satisfying the following harmonic property:

$$\mathbf{f}_l^{(i)} = \frac{\sum_{\mathbf{x}_j \in \mathcal{N}_i} w_{ij} \mathbf{f}_l^{(j)}}{\sum_{\mathbf{x}_j \in \mathcal{N}_i} w_{ij}} \quad (17)$$

Thereby, the Laplacian score desires the feature value at each point could be estimated by the nearest neighbors regression (equivalent to the classical Nadaraya-Waston algorithm [9]), *i.e.*, a weighted average of the values of its neighbors, and the weight of each neighbor is proportional to the proximity.

Therefore both the LKR score and the Laplacian score could be interpreted from a local regression perspective and they both select the features that could keep the local information. One of their main differences is that, rather than LKR score explicitly aiming to estimate the feature value at each point by its neighbors with the kernel ridge regression approach, Laplacian score implicitly performs the estimation by the nearest neighbors regression method.

Furthermore, after taking a closer study on the regression coefficients in (10) and (17), we could note that nearest neighbors regression considers only the distances between \mathbf{x}_i and its neighbors $\mathbf{x}_j \in \mathcal{N}_i$, and ignores the distances between the neighbors; therefore, \mathbf{x}_i may be close to points that are far from each other, resulting in a weighted average

of feature values from two distant and thus likely unrelated points. By contrast, kernel ridge regression considers the distances between pairs of points in the neighbors when deciding how heavily to weigh the influence of relevant neighboring points, which is embodied in the computation for the regression coefficients (*c.f.* (8)). Thus the kernel ridge regression is expected to be more faithful to reveal the local relationship of data than the nearest neighbor regression.

3.2 Connection to other local regression methods

The idea of local regression has been successfully applied recently. The paper in [15] has utilized the local regression error term as a regularizer for the transductive classification, based on the assumption that the real valued class label of each point should be similar to the output of the local regression model, which is trained with its neighbors and their labels. Wu *et al.* [14] propose a local learning clustering algorithm which shares the similar assumption, but minimizes the regression error for the cluster label of each point. In [16], a dimensionality reduction algorithm is proposed based on the kernel regression, it seeks a linear projection satisfying that the projection value of each point can be well estimated based on its neighbors and their projection values. Promising performance reported in [14, 15, 16] all indicates that the local regression approach which minimizes the local estimation error, is effective in exploring the local relationship. In this paper, we perform the regression with the feature values as the model targets, and utilize the local estimation error as a criterion for selecting the features. The kernel ridge regression is adopted in this paper, however, other regression techniques could also be adopted to build the regression model.

4 Experimental results

In this section, we empirically evaluated the performance of LKR score in clustering and classification applications. In all the experiments, the ridge parameter λ was set to be 0.1, and we examined the impact of it on the performance of the algorithm latter in this section. The Radial Basis Function (denoted as RBF) kernel function and the diffusion (denoted as DIF) kernel function [6] were tried in all the experiments.

4.1 Feature selection for face clustering

A subset of the CMU PIE face database was used in this experiment. It contains 68 human subjects of the frontal poses (C27) but under different lighting conditions, with

each subject having 21 faces. We used the cropped images² that have been aligned, normalized and contain only the facial areas, the size of cropped image is 32×32 pixels, with 256 grey levels per pixel. We then scaled the features (pixel values) to $[0,1]$ (divided by 256), and each image was represented by a 1024-dimensional vector.

For a given number c ($c = 5, 10, 30$), data from c classes were randomly selected out of the whole database. This process was repeated 20 times, and the average k -means clustering performance after the feature selection is reported. For the k -means clustering (the number of clusters was set to c), we started from 10 different random initializations and chose the solution with the highest objective function value of the k -means. The unsupervised Laplacian score and LKR score were computed to select features on the same picked c -class data. For both methods, a weighted neighborhood graph was built with the neighborhood size of 5, and the kernel width of the RBF kernel function was set to the same value in the two methods when the RBF kernel function was used in LKR score. Two clustering performance evaluation indexes were used to assess the quality of the selected feature subset for clustering.

Clustering Accuracy

Clustering *Accuracy* describes the match degree between the generated label and the ground-truth class label for each point. Given a data point \mathbf{x}_i , its ground-truth class label y_i provided by the database, and the obtained cluster label c_i for it, the clustering accuracy is computed as follows:

$$Accuracy = \frac{\sum_{i=1}^n \delta(y_i, \text{map}(c_i))}{n} \quad (18)$$

where n is the number of the data set, and $\delta(x, y)$ is the delta function that equals 1 if $x = y$ and equals 0 otherwise. The $\text{map}(\cdot)$ is a permutation mapping function that maps each cluster index c_i to a true class label. The clustering accuracy is defined as the minimal classification accuracy among all possible permutation mappings. This optimal matching can be found with the Kuhn-Munkres algorithm [10], which is devised for obtaining the maximal weighted matching of a bipartite graph.

Normalized Mutual Information

The *Normalized Mutual Information* (NMI) is widely used to measure the similarity between the produced groupings and the true groupings indicated by the class labels. For two random variable \mathbf{X} and \mathbf{Y} , the NMI is defined as follows:

$$NMI(\mathbf{X}, \mathbf{Y}) = \frac{I(\mathbf{X}, \mathbf{Y})}{\sqrt{H(\mathbf{X})H(\mathbf{Y})}} \quad (19)$$

²The data is obtained from <http://www.cs.uiuc.edu/homes/dengcai2/Data/data.html>

Table 1. Results of the 20-fold runs on the PIE C27 database

		Accuracy				
c	no. of features	24	124	224	324	1024
5	LKR(RBF)	0.787	0.868	0.909	0.832	0.490
	LKR(DIF)	0.508	0.824	0.817	0.782	0.490
	Laplacian	0.707	0.755	0.832	0.795	0.490
10	LKR(RBF)	0.634	0.753	0.784	0.702	0.403
	LKR(DIF)	0.403	0.615	0.679	0.702	0.403
	Laplacian	0.590	0.702	0.730	0.672	0.403
30	LKR(RBF)	0.563	0.598	0.573	0.539	0.347
	LKR(DIF)	0.415	0.548	0.562	0.539	0.347
	Laplacian	0.520	0.573	0.537	0.539	0.347
		NMI				
c	no. of features	24	124	224	324	1024
5	LKR(RBF)	0.824	0.867	0.895	0.772	0.480
	LKR(DIF)	0.669	0.801	0.825	0.778	0.480
	Laplacian	0.748	0.832	0.825	0.747	0.480
10	LKR(RBF)	0.782	0.841	0.843	0.780	0.548
	LKR(DIF)	0.677	0.806	0.804	0.790	0.548
	Laplacian	0.758	0.819	0.818	0.761	0.548
30	LKR(RBF)	0.797	0.820	0.778	0.755	0.616
	LKR(DIF)	0.740	0.796	0.796	0.763	0.616
	Laplacian	0.770	0.789	0.767	0.750	0.616

where $I(\mathbf{X}, \mathbf{Y})$ is the mutual information between \mathbf{X} and \mathbf{Y} , while $H(\mathbf{X})$ and $H(\mathbf{Y})$ are the entropies of \mathbf{X} and \mathbf{Y} , respectively. Given the obtained clustering result, the NMI is estimated by:

$$NMI = \frac{\sum_{i=1}^c \sum_{j \in \mathcal{C}'} n_{i,j} \log\left(\frac{n \cdot n_{i,j}}{n_i \tilde{n}_j}\right)}{\sqrt{(\sum_{i=1}^c n_i \log \frac{n_i}{n})(\sum_{j \in \mathcal{C}'} \tilde{n}_j \log \frac{\tilde{n}_j}{n})}} \quad (20)$$

where n is the number of samples in the data set, n_i denotes the number of data contained in the cluster \mathcal{C}_i ($1 \leq i \leq c$), \tilde{n}_j is the number of data belong to the j -th class ($j \in \mathcal{C}'$ the set of selected ground truth classes), and $n_{i,j}$ denotes the number of data that are in the intersection between the cluster \mathcal{C}_i and the j -th class. The NMI ranges from 0 to 1, and the larger the value, the more similar the groupings by clustering and those by the true class labels.

The results over different no. of selected features are summarized in Table 1. We could observe that the performance of clustering on the feature subset selected either by the LKR score or the Laplacian score improves that using all the features, indicating that there are indeed only a fraction of the 1024 features that are useful for the partitioning. We can also see that LKR score using RBF kernel works consistently better than Laplacian score, in terms of both evaluation metrics, and the LKR score using the diffusion kernel sometimes slightly outperforms the others.

4.2 Feature selection for classification

In this subsection, we investigated the performance of our proposed LKR score feature selection for the face

recognition and the cancer microarrays classification. For the face recognition applications, we reported the results on the PIE C27 dataset used above and the Yale-B10p dataset. The Yale-B10p is a subset of the Extended Yale-B database, it contains 10 human subjects under the frontal pose and 64 different illuminations. The cropped images³ were used, and the size is 32×32 pixels, with 256 grey levels per pixel. We then scaled the features to $[0,1]$, and represented each image with a 1024-dimensional vector. For the cancer microarrays classification, we used two public datasets. The 1000-dimensional lung cancer data⁴ includes 4 known classes: 139 adenocarcinomas (AD), 21 squamous cell carcinomas (SQ), 20 carcinoids (COID), and 17 normal lung (NL) [8]. The colon cancer data⁵ contains 2000 genes over 62 samples from 2 classes of colon-cancer patients: 40 normal healthy samples and 22 tumor samples. The publicly available data were pre-processed following the similar methods in [3]. First we thresholded the data set with floor of 1 and ceiling of 16000. Then we filtered out genes with $max/min \leq 5$ or $(max - min) \leq 500$, where max and min are the maximum and minimum expression values of a gene. After a base 10 logarithmic transform, each sample was standardized to zero mean and unit variance across genes. The characteristics of datasets eventually used in this subsection are summarized in Table 2.

Table 2. Characteristics of datasets used for classification

Dataset	samples	classes	features
PIEC27	1428	68	1024
YaleB10p	640	10	1024
Lung cancer	197	4	419
Colon cancer	62	2	1224

For each dataset, two-thirds of the whole samples in each class were randomly selected to form the training set, and the test set consisted of the remaining samples. The feature selection was performed on the training set, and we recorded the 1-nearest neighbor (1NN) classification accuracy on the test set, using the selected features. This splitting was repeated for 20 times for face recognition experiments, and 100 times for cancer analysis, since the cancer datasets have much less samples. The average classification accuracy over these splits were used to evaluate the quality of selected features.

To appreciate the strength of our approach to identify discriminative features, we firstly ran the LKR score algorithm on these problems without the class labels in the training phase (unsupervised version), and then studied the

³The data is obtained from <http://www.cs.uiuc.edu/homes/dengcai2/Data/data.html>

⁴<http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi>

⁵<http://microarray.princeton.edu/oncology/affydata/index.html>

case with the class labels involved in the training phase (supervised version). Note that supervised feature selection is generally expected to perform better than unsupervised methods.

Study of unsupervised feature selection

The unsupervised Laplacian score was compared with our method again in this study. For both the LKR score and the Laplacian score, a 10-nearest neighborhood graph was constructed. The accuracy vs. no. of selected features is plotted in the first column of Figure 1, and the average accuracy on each data set over all numbers of selected features (see Figure 1) is shown in Table 3. As can be seen from Figure 1 and Table 3, LKR score using RBF kernel function works much better than Laplacian score on the face image datasets, and is comparable to Laplacian score on the two cancer datasets. LKR score using the diffusion kernel function works robustly on all datasets in this study. In addition, we could even observe that the unsupervised version of LKR score using diffusion kernel function is only slightly inferior to the baseline supervised feature selection method compared in the supervised feature selection experiments below.

Study of supervised feature selection

The popular Fisher score is shown to be equivalent to the supervised extension of the Laplacian score [5], thus was compared as a baseline in the supervised case. The second column of Figure 1 shows the 4 plots of accuracy vs. no. of selected features. We could observe that LKR score using RBF kernel function performs the best on most datasets except the Colon cancer dataset, which has much less samples than other datasets. Particularly, applying LKR score with RBF function could improve the performance of Laplacian score by a big margin on the PIE C27 and Lung cancer dataset when less than 40 features were selected. The average accuracy on each data set over all numbers of selected features (see Figure 1) is shown in Table 3. The experimental results indicate that for classifying sufficiently large datasets, LKR score using RBF kernel function is a favorable candidate for performing the supervised feature selection.

Study of the effect of λ

Compared to the Laplacian score, the LKR score has an extra parameter λ , which controls the smoothness of the kernel ridge regression estimator, and makes sure that the matrix in(7) is invertible. It was set to 0.1 in the previous experiments, now we tried some different reasonable λ 's in a broader interval, and examined their effects on the feature

Table 3. The average performance over all numbers of selected features

Unsupervised feature selection				
Method	PIE C27	YaleB10p	Lung	Colon
LKR(RBF)	0.8844	0.6311	0.9113	0.7385
LKR(DIF)	0.7144	0.7195	0.9145	0.8066
Laplacian	0.7510	0.3562	0.9100	0.7396
Supervised feature selection				
Method	PIE C27	YaleB10p	Lung	Colon
LKR(RBF)	0.8266	0.7919	0.9177	0.7614
LKR(DIF)	0.6946	0.7461	0.8901	0.8197
Fisher	0.7349	0.7408	0.8872	0.8199

quality. The third column in Figure 1 shows the results in which the performance of LKR score is plotted as a function of λ ($\lambda \in [0.05, 10]$), with the number of selected features fixed to 50 for all datasets. For the PIE C27, the Lung cancer and the Colon cancer datasets, the difference between the highest accuracy and the lowest one on each curve is no more than 0.05; for the YaleB10p dataset, the largest differences occur on the curves for the unsupervised case, but they are less than 0.2, and LKR score still outperforms the unsupervised baseline method (i.e., Laplacian score) on this dataset when λ is set other values instead of 0.1. Thus, the selection of λ may not be a very crucial problem in LKR score.

5 Conclusion

In this paper, we have proposed a new feature selection method based on the local kernel regression and the neighborhood graph. After obtaining the neighborhood graph, the features that could hold the structure of local patch on the graph and have enough representative power are selected. And by adopting different definitions of the neighborhood, the proposed feature selection algorithm can be performed either in the unsupervised or the supervised manner. The experimental results have shown that its effectiveness in identifying discriminative features that contribute to revealing the underlying data structures. Our future work is to explore the proposed LKR score in the semi-supervised problems, where the class labels are partially given.

References

[1] A. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2):245–271, 1997.
 [2] F. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997.

[3] S. Dudoit, Y. Yang, M. Callow, and T. Speed. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12(1):111–139, 2002.
 [4] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3(7-8):1157–1182, 2003.
 [5] X. He, D. Cai, and P. Niyogi. Laplacian score for feature selection. *Advances in Neural Information Processing Systems*, 18, 2005.
 [6] R. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete structures. In *Proceedings of International Conference on Machine Learning*, 2002.
 [7] H. Liu and L. Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, 17(4):491–502, 2005.
 [8] S. Monti, P. Tamayo, J. Mesirov, and T. Golub. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, 52(1):91–118, 2003.
 [9] E. A. Nadaraya. *Nonparametric Estimation of Probability Densities and Regression Curves*. Kluwer Academic, 1989.
 [10] C. Papadimitriou and K. Steiglitz. *Combinatorial Optimization: Algorithm and Complexity*. Dover, 1998.
 [11] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.
 [12] A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-posed Problems*. John Wiley, New York, 1977.
 [13] L. Wolf and A. Shashua. Feature selection for unsupervised and supervised inference: The emergence of sparsity in a weight-based approach. *Journal of Machine Learning Research*, 6:1855–1887, 2005.
 [14] M. Wu and B. Schölkopf. A local learning approach for clustering. *Advances in Neural Information Processing Systems*, 19, 2007.
 [15] M. Wu and B. Schölkopf. Transductive classification via local learning regularization. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*, 2007.
 [16] M. Wu, K. Yu, S. Yu, and B. Schölkopf. Local learning projections. In *Proceedings of the 24th International Conference on Machine Learning*, pages 1039–1046, 2007.
 [17] Z. Zhao and H. Liu. Spectral feature selection for supervised and unsupervised learning. In *Proceedings of International Conference on Machine Learning*, pages 1151–1158, 2007.

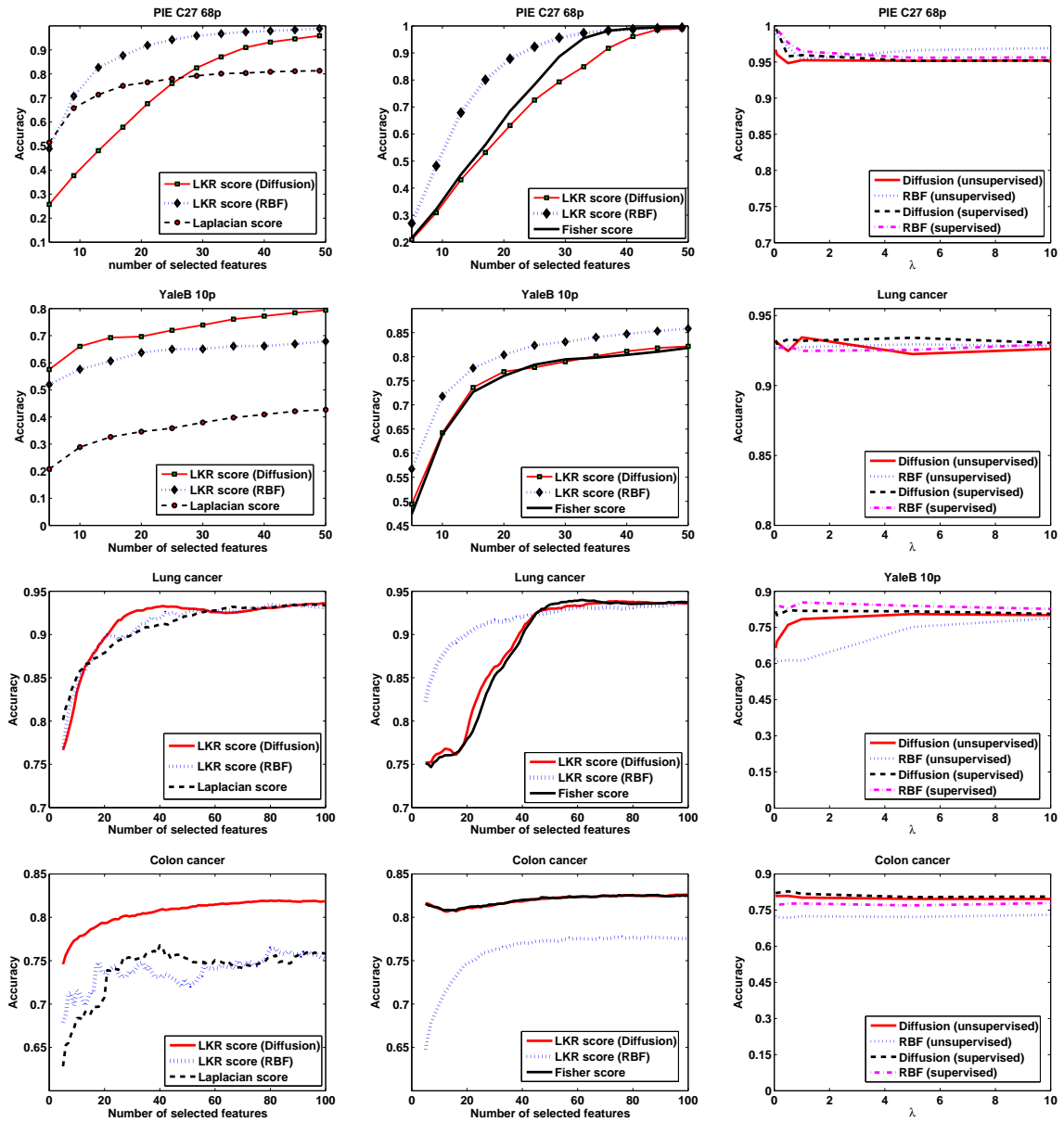


Figure 1. The comparison of performance. Figures in each row describes the results on each dataset. The figures in the first column are for the unsupervised feature selection, and the ones in the second column are for supervised feature selection, the word in the parentheses stands for the kernel function used. The figures in the third column plot the effect of λ on the performance of LKR score, and the word in the parentheses stands for the unsupervised or supervised manner in which the LKR score performs.

A Survey of Community Mining in Complex Networks

Dan Zhang

Abstract

Community structure is a significant topological property of complex networks. It can be found in many real-world complex networks, such as social groups with similar backgrounds or interests, web pages containing similar topics, modules in metabolic or cellular networks, image segments with the similar color. Mining these communities is crucial. It has attracted extensive interests of researchers in recent years. There are many methods developed in various areas. In this paper, we give an overview of these methods including both traditional methods and some new ones. In particular, we focus on the two latest Autonomy-Oriented Computing (AOC) based methods and compare their performances in detail.

1 Introduction

Many complex systems in the real world can be described as complex networks, i.e., graphs composed of nodes and links between them [10, 20]. One typical characteristic of complex networks is community structure, which is always considered to be a group of nodes densely connected inside and sparsely connected outside. Community structures exist in a variety of real-world complex networks, such as social groups with similar backgrounds or interests, web pages containing related topics, modules in metabolic or cellular networks, image segments with the similar color. Mining these communities is crucial, it makes us know more about the topological structures of the networks, and allows us to utilize the resulting communities or mine information in a more efficient way. For example, detecting communities in the World Wide Web networks can help us to make the search engines more powerful with automatic classification.

With respect to a concrete problem, it is crucial to formulate the problem in generic complex networks. Two steps are need to be implemented: define the community structure and design the methods to mine these community structures.

The remainder of this paper is organized as follows: Section 2 discusses two general definitions of community structures. Section 3 summarizes some traditional community

mining methods. In Section 4, we further introduce the two latest AOC-based methods in detail. Section 5 compares the performances by these two methods in both clustering quality and time complexity. Finally, we conclude the paper by highlighting the main work.

2 Definition of community structures

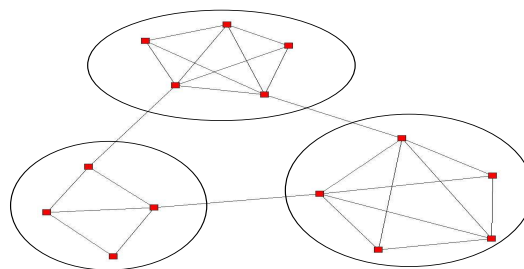


Figure 1. An example of subgraph-based community structure in a network. There are three communities labeled by ellipses containing 5, 4, 5 nodes respectively. The links within these communities are dense but between them are sparse.

In terms of community structures, there is no formal definition up to now. Literatures show that different definitions should be given according to different situations.

Many literatures adopt a subgraph-based definition. Which is, community is a subgraph within which vertex-vertex connections are dense but between connections are sparse (as shown in Fig.1). This description is originally proposed by Girvan and Newman in [12], and then applied in a series of papers [11, 8, 12, 13, 14, 5, 16, 19]. In particular, Radicchi et al have given a quantitative formulas of this definition in [4].

However, sometimes communities can't be defined in this way. A specific example is shown in Fig.2. This is a bipartite network containing two types of nodes. The upside square nodes represent customers, which connect to

different bank staffers downside. Since there is no single connection between each other in the customers community and bank staff community, they can't be represented by the subgraph-based definition any more. In this case, community is always considered to be a set of nodes that have the same neighbors or preferences in their connections.

Caldarelli have given a detailed illustration about the definitions of communities in [2]. He discriminated community from other synonymous concepts such as class, cluster, clique, subgraph respectively. As Caldarelli pointed out, "a cluster is associated to a community of some kind, some communities do not correspond to clustered subgraphs".

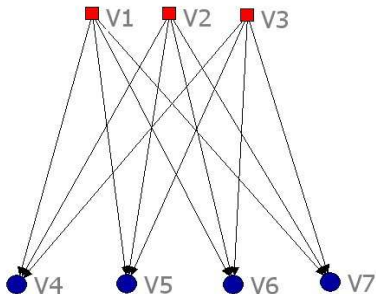


Figure 2. A simple bipartite network. The network contains two types of nodes with three customers upside (square nodes V1, V2, V3) connecting four different bank staffers downside (circle nodes V4, V5, V6, V7), but with no single connection between each other.

Therefore, different definitions should be given according to different situations. They mainly depend on the attributes of the nodes and the topological properties of the given networks. For general research purposes, most researchers prefer to the subgraph-based definition. Generally, communities are defined as subgraphs with high link densities in unweighted, undirected and unsigned networks. However, in the cases of weighted, directed and signed networks, it needs to modify the definitions somewhat. For the weighted and signed cases, communities should be defined as subgraphs in which positive links are dense and negative links are sparse. For the directed case, take web communities for example. Flake et al [5] defined them as graphs, in which each vertex has more connections than out of it, i.e., the strong sense definition mentioned in [4]. In sum-

mary, community structures should define according to different situations to coincide the convenience of one's study. Generally, it is reasonable and useful to adopt the sub-graph based definition.

3 Traditional community mining methods

The study of community mining has a long history due to its applications in many fields. It is closely related to graph partitioning in graph theory and computer science, hierarchical clustering in sociology. A large number of community mining methods have been developed in various areas recently. Generally, these traditional methods can be summarized as two families along their different backgrounds:

- I. Hierarchical methods, which are a series of sociology-based methods that select suitable groups by processing special nodes and edges present in the networks.
- II. Spectral methods, which are graph theory-based methods in mathematics that select communities according to the properties of the eigenvectors from the adjacency matrix.

3.1 Hierarchical methods

Hierarchical methods include two representative methods: agglomerative method [15] and divisive method [8]. The main idea of hierarchical methods is mapping a network to a tree with nested hierarchical structures by adding or removing edges.

In the agglomerative method, the communities are generated by adding edges. Starting from the separate nodes with no links, one firstly calculates the similarities between every pair of nodes in the network, and then adds edges between them one by one in the order from the strongest to the weakest similarities. By virtue of this agglomerative method, large subgraphs emerge increasingly. Then communities composed of these nested subgraphs are detected. To implement the agglomerative method, several formulas of measuring the similarities between node pairs can be chosen. A summarization of various similarity measurements can be found in [2]. All of these quantities give reasonable results for mining communities. Nevertheless, some single peripheral nodes always can't be correctly separated by this method. Since single node often remains isolated from the network when the node connects to the rest nodes by only a single edge.

To sidestep the shortcomings of the agglomerative method, Girvan and Newman proposed a reverse method by removing the most crucial edges. This division method is referred as the so-called GN method for abbreviation. The GN method focuses on the boundaries of communities rather than their cores as the agglomerative method does.

Girvan and Newman first defined a quantity called edge betweenness. It aims to measure the importance of an edge in the information propagation in a network, and is a generalized notion of the traditional vertex betweenness centrality. Analogous to the vertex betweenness centrality, the edge betweenness is defined as the fraction of shortest paths running through this edge among all the shortest paths between node pairs. The higher the edge betweenness is, the more important the edge is. Thereby, this kind of large betweenness edges quite accord the role of links between communities. Assume removing these edges with large betweenness, possible communities can be naturally divided. However, the GN method has a high time complexity $O(n^3)$ (n is the total number of nodes in a network), it is always unavailable for the large scale data sets. To improve the running speed and the performances, lots of alternative GN methods have been proposed such as fast GN algorithm [13], self-contained GN algorithm [4] and Tyler algorithm [14].

3.2 Spectral methods

Spectral methods [2, 20] belong to a class of mathematical methods. The community structures are determined by the eigenvalues and eigenvectors of suitable functions of the adjacency matrix $A(i,j)$.

The basic spectral method is the spectral bisection method [1] proposed by Pothen et al in 1990. It is based on the processing of the Laplace matrix L , which is defined using the adjacency matrix. If a link exists between vertex i and j , the element $L_{ij} = -1$, and the diagonal elements contain degree of node i , i.e, $L_{ii} = d_i$. Empirical studies show that the adjacency matrix is made of distinct blocks when the communities are made by separate subgraphs. It means that every block represents for a particular subgraph. Suppose there are only two communities in a network, then the corresponding Laplace matrix contains two diagonal matrix blocks. It is known that for any real symmetric matrix, its nontrivial eigenvectors are always orthogonal, thus these eigenvectors definitely have positive elements as well as negative ones. Therefore, the network can be cut into two communities according to the nodes corresponding to the positive elements and negative elements respectively.

The traditional spectral methods just allow us to divide networks into two communities, this bisection function is far from the real applications. Additionally, this method runs very slow, since calculating all the eigenvalues of a $n \times n$ matrix takes $O(n^3)$ (n is the total number of nodes in a network) time complexity. As a result, many improved spectral methods have been proposed to heighten the precision and cut down the time complexity, such as the multi-dimensional spectral method introduced by Donetti in [9]. For more details of this kind of methods, one may see [3].

3.3 Other methods

Apart from the above two traditional families of community mining methods, a great deal of new ones have been developed in recent years. Among those methods, some are oriented to solve special problems.

Clique percolation method (CPM) [7] was proposed by Palla et al in 2005. It aims at analyzing overlapping communities. No matter hierarchical methods or spectral methods, both of them are fit for the cases that the communities are independent with each other. However, most real-world networks have no absolute independent community structures as ideal situations, instead they are always link or overlap to each other. CPM is designed to be becoming to this situation. CPM defines k -clique community as a union of k -cliques that can be reached through a series of adjacent k -cliques, where k -clique refers to fully connected subgraph with k nodes, and adjacent k -cliques refer to the two k -cliques sharing $k-1$ nodes. This physical method contains two steps: first locate all cliques of the network and then performs a standard component analysis of the clique-clique overlap matrix to discover the k -clique-communities. Palla et al pointed the time complexity of the CPM is about $\alpha n^{\beta \ln(n)}$ (n is the total number of nodes in a network, α, β are constants) from the empirical studies.

For other special purposes, various methods were proposed as well. Some researchers tend to pay their attention to signed networks. For instance, Yang et al [17] recently published a method to mine communities in signed network with both positive within-community and negative between-community links are dense. This method recognizes every vertex as an agent, so that it can visit to other nodes to find its own community members. The agent makes the selection by choosing those nodes with high local aggregated transition probabilities, which can be calculated by iterative operation on the adjacency matrix of the network. For detecting communities in web pages, the most representative method is maximum flow communities (MFC) method [5] proposed by Flake et al.

3.4 Distributed methods

So far, all the community mining methods mentioned above are concerned with the centralized networks. In other works, they are available only when the global topological structures are provided. These methods require us to collect the information about all the nodes and links beforehand. Nevertheless, it is always too hard to implement it in the cases of the real-world networks. For instance, in the World Wide Web network, it is almost impossible for us to get all the web pages together, since there are so many pages existed in the world. In addition, these web pages may dynamically update every day. For such distributed, dynami-

cally evolved networks, the mentioned traditional methods are unable to deal with them anymore. In most cases, we are only permitted to get some limited information about the networks. Consequently, new methods based on the local views are extraordinarily needed.

Yang and Liu [19] have proposed such a kind of method recently, which based on the Autonomy-Oriented Computing (AOC). AOC [18] is an efficient computing paradigm presented by Liu et al in 2001, which makes the entities in a complex system self-organize by themselves and model autonomy to meet some computing requirements. The proposed AOC-based method is based on the local interactions between some node pairs, where the interactive rules obey some certain conditions that are defined to be propitious for communities emerging. Because every node only transacts with its current local neighbors, there is no need to require all the nodes participating in. In other words, it is available even some nodes are unprovided. Yang and Liu showed that his method has a good performance in [19]. We will review the detailed algorithms of this method in the following sections.

Based on the node-node interactions, Frey and Dueck et al has presented another powerful method named Affinity-propagation (AP) method in [6]. This method invites the self-organizations of the nodes as well, in which the interactions are called messages passing. Thereby, the Affinity-propagation method is recognized as a kind of AOC-based method either. As [6] indicates, this method obtains good results when applied in the real-world data.

In the following sections, we will introduce these two AOC-based methods in detail and give some comparison analysis. In order to avoid misunderstanding, we rename the AOC-based method proposed by Yang and Liu in [19] as AOC-YL method in the following parts.

4 Autonomy-Oriented Computing based methods

4.1 AOC-YL method

The AOC-YL method was proposed by Yang and Liu recently [19]. It concerns with the distributed and dynamic networks rather than the traditional considered centralized and static ones. The AOC-YL method is based on the local views of the nodes. It dynamically detects the communities via interactions between nodes and their neighborhood nodes.

More amply, we represent each node as an agent, i.e., if there are N nodes in a network there are N agents in the model. To measure an agent's state, agent view is defined as $agent_p = (V_p, E_p)$. Thereinto, V_p represents the set of nodes controlled by the agent p (A_p for abbreviation), and E_p is the set of edges going out from these nodes. In

other words, A_p 's view is a set of links between A_p and its neighbors with weights and directions information. To find communities is to update the agents view based on the local information until the view cannot be changed any more.

Here we summarize the AOC-YL algorithm in the Table.1 as follows. In which, $S(A_p, A_q)$ represents the similarity between agent p and agent q . " \leftarrow " denotes "assign to".

Table 1. AOC-YL algorithm

Input: Adjacency matrix of a network.
Output: The label vector of the nodes belonging to the detected communities.

for each agent p **do**
Step 1: Evaluating the agents view.
 Calculate the similarities between agent p and its neighbors.
Step 2: Shrinking/enlarge agent view.
 Calculate the threshold T . T is composed by a linear function of the current mean m_p and standard deviation σ_p , $T = \omega_1(m_p + \omega_2\sigma_p)$ (ω_1 and ω_2 are constants between 0 and 1).
 if $S(A_p, A_q) < T$, set $S(A_p, A_q) \leftarrow 0$.
 (The friendship is broken up, drive A_q out from A_p 's view.)
 else $S(A_p, A_i) \leftarrow S(A_p, A_i)$.
 (Keep the friendship unchanged.)
 Calculate $S(A_q, A_i)$ and $S(A_p, A_i)$, thereinto, A_i is A_q 's neighbor.
 if $S(A_p, A_i) < S(A_p, A_q) \cdot S(A_q, A_i)$,
 set $S(A_p, A_i) \leftarrow S(A_p, A_q) \cdot S(A_q, A_i)$;
 (Make friends with friend's friend, add A_i into A_p 's view.)
 else, $S(A_p, A_i) \leftarrow S(A_p, A_i)$.
 (keep the friendship unchanged.)
 endif
endif
Step 3: Re-calculate the current threshold T and re-shrink the agent view as **Step 2**.
Step 4: Balancing.
 if there is both links from A_p and A_q ,
 $S(A_p, A_q) \leftarrow (S(A_p, A_q) + S(A_q, A_p))/2$;
 else $S(A_p, A_q) \leftarrow (S(A_p, A_q))/2$.
 endif
endfor

We have realized this agent-based AOC-based method on a set of benchmark data. In the following section we demonstrate their performances in detail.

4.2 Affinity-propagation method

The Affinity-propagation method was proposed by Frey and Dueck in [6]. This method identifies exemplars (namely those data points can be chosen as the centers such that the sum of squared errors between data points and their nearest centers is small) by recursively sending real-valued messages between pairs of data points. By passing the real-valued messages between data points pairs, each data point can calculate its current affinity to determine whether choose another data point as its exemplar. After enough iterations, each data point may find its own exemplar and the corresponding communities can automatically emerge.

As Frey and Dueck pointed out, the messages are described by two defined quantities: responsibility and availability.

- Responsibility $r(i, k)$:

$$s(i, k) - \max_{k \neq k'} (a(i, k') + s(i, k')) \quad (1)$$

Where $s(i, k)$ is the similarity between data point i and k , $a(i, k)$ is the availability from data point k to i defined as below. The responsibility $r(i, k)$ is from node i to candidate exemplar k , to evaluate how well suited is exemplar for data point i , compared to all other possible exemplars.

- Availability $a(i, k)$:

$$s(i, k) - \min\{0, r(k, k) + \sum_{i' \neq i, k} \max(0, r(i', k))\} \quad (2)$$

$$a(k, k) = \sum_{i' \neq i} \max(0, r(i', k))$$

Where $r(i, k)$ is the responsibility from data point i to k . The availability $a(i, k)$ is from candidate exemplar k to data point i , to evaluate how appropriate is candidate k as exemplar for data point i , taking support from other data points into account.

Given a network, we have to calculate the similarities between the nodes first, then update each node's responsibility and availability according to the formulas above. For node i , the sum of the current responsibility and availability $a(i, k) + r(i, k)$ denotes the affinity that node i chooses node k as its exemplar. Therefore, the value of k that maximizes $a(i, k) + r(i, k)$ will be the exemplar of node i . The algorithm can be summarized as the following five steps:

This method can select the cluster exemplars automatically, it has a better performance on some real-world data compared with the k-centers clustering algorithm. For more

Table 2. Affinity-propagation algorithm

Input: Similarity matrix S .

Output: The label vector of the nodes belonging to the detected communities.

Step 1: Update the responsibility matrix R , where

$$R_{i,k} = r(i, k). \text{ Calculate } r(i, k) \text{ as given in Eq.1.}$$

Step 2: Update the availability matrix A , where

$$A_{i,k} = a(i, k). \text{ Calculate } a(i, k) \text{ as given in Eq.2.}$$

Step 3: Find the maximum elements in the diagonal of $R + A$ to determine the exemplars.

details, one may see the [6]. In this paper, we have realized the Affinity-propagation method on some computer-generated data. Fig.3 shows the clustering results on 200 random generated data points.

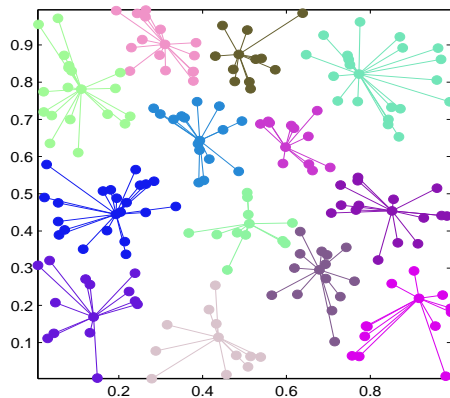


Figure 3. An illustration of the results using the Affinity-propagation algorithm on a network with 200 data points randomly generated. We see that there are total 13 star-networks have emerged, different color of star-network represents different communities generated. The central data point of each star-network represents the exemplar that found.

5 Performances of the two AOC-based methods

We have realized the above two AOC-based methods on the random networks generated by the computer. Comparisons of their performances are given in terms of both the clustering quality and running speed.

Table 3. The number of converged communities by the Affinity-propagation method and the AOC-YL method under four random networks G1, G2, G3 and G4 generated above.

Different random networks	G1	G2	G3	G4
Affinity-propagation method	4	6	8	90
AOC-YL method	4	4	4	4

5.1 Experimental design

We operate the two methods on a series of random networks generated by computers, which is commonly used for testing community mining methods [12, 16, 19]. We record a random network as $G(c,n,k,p)$, where c is the number of communities, n is the number of nodes in each community, k is the degree of each node and p is the probability of edges in the same community. In this paper, four random networks are generated for evaluation, they correspond to $G1(4,10,10,0.6)$, $G2(4,25,20,0.6)$, $G3(4,64,30,0.5)$ and $G4(4,200,100,0.6)$ respectively.

In our experiments, we will examine the performances of these two AOC-based methods. Two items are evaluated: clustering quality and running speed. In addition, we will demonstrate some visible results.

In the AOC-YL method, two parameters ω_1, ω_2 as mentioned in Table.1 are assigned to $\omega_1 = 0.4, \omega_2 = 0.2$. In the Affinity-propagation method, we set the damping factor as $\lambda = 0.2$. Each method on every network will be executed for several times, snapshots with the best results of both the two methods will be recorded and plotted.

5.2 Experimental results

Table.3 shows the number of communities divided after convergence under the four networks G1, G2, G3 and G4 defined above with total 40, 100, 256 and 800 nodes respectively. Compared with the real correct community number $c = 4$, we find that the AOC-YL method clustered the quantity of communities correctly, while the Affinity-propagation method failed on three larger size generated random networks G2, G3, G4.

In order to further evaluate the performances of the two methods, we engaged modularity as a measurement. This quantity is presented by Newman and Girvan [12] to evaluate the quality of the division of a network. Suppose there are K communities separated finally, E is a symmetric matrix where E_{ij} is the proportion of all edges connecting community i and community j , E_{ii} is the fraction within

the community i . Then the modularity is defined as:

$$M = TrE - \|E^2\| \tag{3}$$

where $\|E\|$ indicates the sum of the elements of the matrix E . As reported by Newman and Girvan, the higher value of modularity indicates better division of the networks.

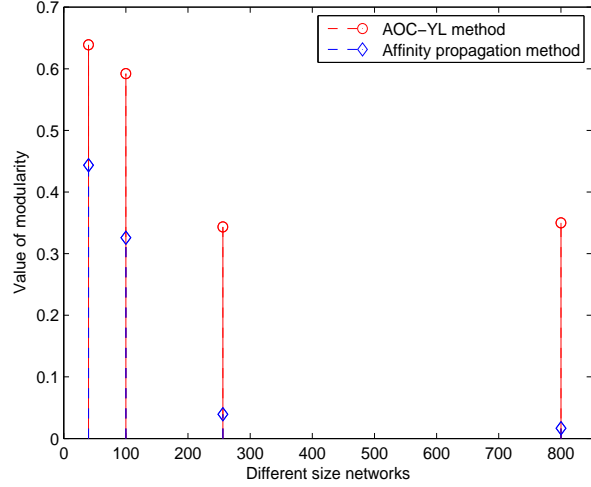


Figure 4. Modularity of the two methods.

We have evaluated this quantity as given in Eq.3. Fig.4 demonstrates the values obtained by the two AOC-based methods versus the network size. The up red line represents the modularity by the AOC-YL method, the values are located as 0.6389, 0.5921, 0.3434, 0.35 respectively. Comparatively, the Affinity-propagation method gets more lower values with 0.4435, 0.3259, 0.0395 and 0.0167 respectively. That is to say, in terms of the division performance, AOC-YL method does better than the Affinity-propagation method.

More details can be seen in Fig.6. It demonstrates the converged graphs of a random network including 40 nodes as shown in Fig.5. As Fig.6(a) shows, AOC-YL method let the nodes interact and evolve to four communities correctly that each community contains the corresponding 10 nodes. In Fig.6(b), although there are still four communities emerged by the Affinity-propagation method, however, we see the central four nodes have been misclassified.

Fig.7 demonstrates the adjacency matrix of the random network with 40 nodes by the two AOC-based methods, on the left is the initial situation and on the right is the situation after convergence. Fig.8 demonstrates the results in case of 256 nodes. The black points indicate that there is a link exist between the nodes in the corresponding positions, contrarily, the white node indicates none. If the four communities are correctly detected, the adjacency matrix will be converged to four black blocks as shown in

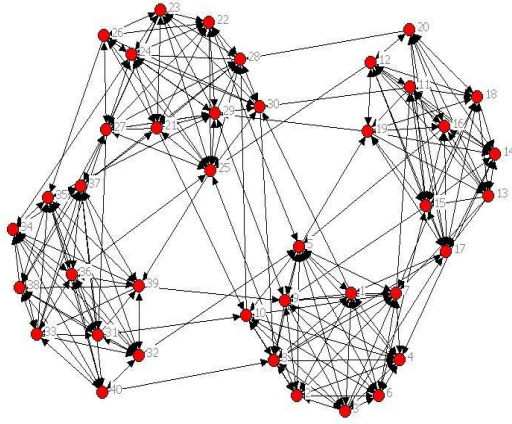
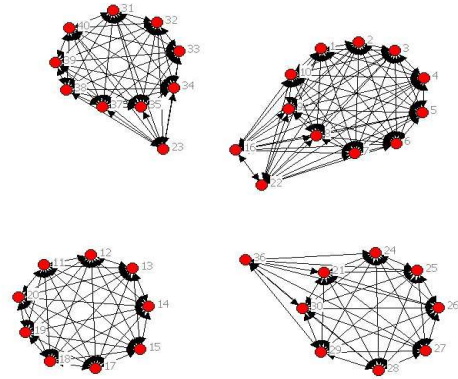
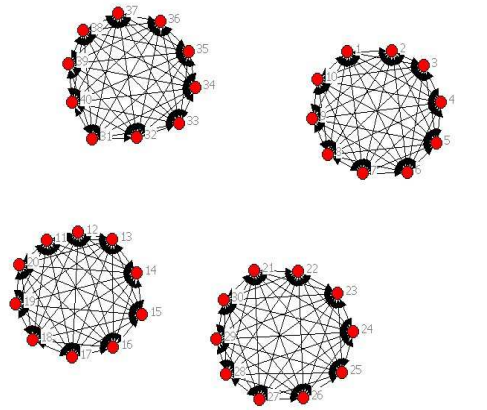


Figure 5. The initial random network $G1(4,10,10,0.6)$.

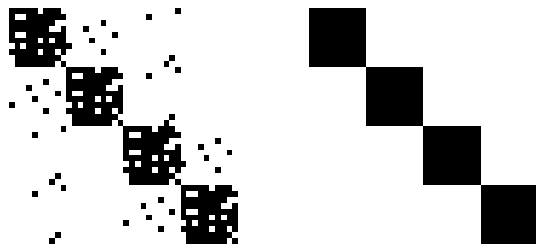
Fig.7(a) and Fig.8(a). Those black blocks indicate the complete subgraphs with high link densities, i.e., communities. The AOC-YL method gives good performances. Comparatively, the adjacency matrix obtained by the Affinity-propagation method demonstrate incomplete or piecemeal blocks rather than complete ones, as shown in Fig.7(b) and Fig.8(b). These phenomena indicate that the communities can't be correctly separated in this case. As Table.3 indicates earlier that there are 90 communities obtained by the Affinity-propagation method, it is not difficult to understand the results shown in Fig.8(b). It is also easy to understand why the modularity of Affinity-propagation method is much lower than the AOC-YL method. In other words, the AOC-YL method does better than the Affinity-propagation method in the aspect of division quality.

However, it always not easy for an algorithm possessing both good performance and low time complexity. Though the AOC-YL method has a better performance of division quality than the Affinity-propagation method, its running speed is much slower than the latter one. As shown in Table.4, the time comparison is significant. Why the Affinity-propagation algorithm runs much faster than the AOC-YL method? In terms of the Affinity-propagation algorithm introduced in Table.2, we see that the time complexity of the step 1 and 2 are $O(n^2)$ and the step 3 is $O(n)$ (n is the total number of nodes in a network). Thus the total time complexity is $O(n^2)$, however it seems higher than $O(n^2/k^4)$ reported in [19] of the AOC-YL method. But why the Affinity-propagation algorithm runs much faster than the AOC-YL method? We find that in each iteration step, the AOC-YL method cost more time to calculate more

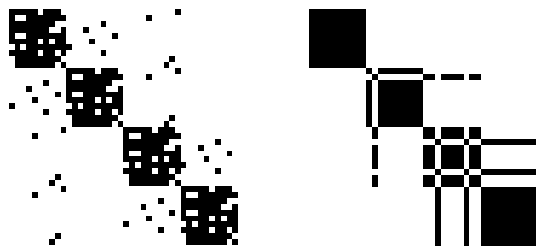


(b)

Figure 6. The converged communities of $G1(4,10,10,0.6)$. (a) Four communities are generated by the AOC-YL algorithm and all are correctly separated. (b) Four communities are generated by the Affinity-propagation method, but four nodes in the center are wrong separated.

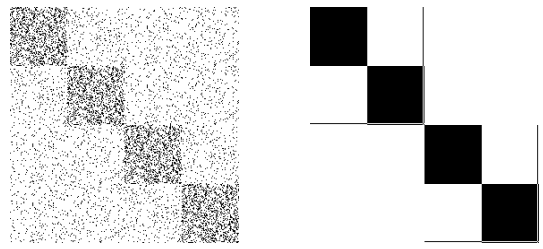


(a)

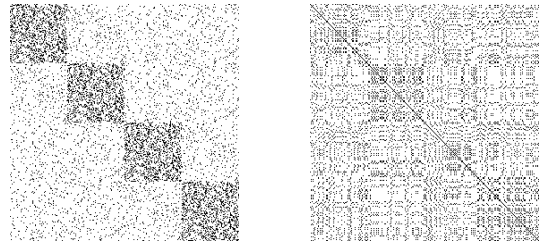


(b)

Figure 7. The converged adjacency matrix of $G1(4,10,10,0.6)$. The black points indicate that there is a link exist between the nodes in the corresponding positions, contrarily, the white node indicates no link. (a) Four communities correctly generated by the AOC-YL algorithm, each community contains 10 nodes. (b) Four communities generated by the Affinity-propagation algorithm, however with some nodes misclassified.



(a)



(b)

Figure 8. The converged adjacency matrix of $G3(4,64,30,0.5)$. (a) Four communities correctly generated by the AOC-YL algorithm, each community contains 64 nodes. (b) Many communities generated by the Affinity-propagation algorithm, with rare nodes correctly separated.

Table 4. CPU time comparison (seconds)

Different random networks	G1	G2	G3	G4
Affinity-propagation method	0.12	0.39	3.29	83.28
AOC-YL method	0.40	2.25	48.35	352.9

quantities such as the similarities, means and standard deviations, while the the Affinity-propagation method only needs to select the extremum values as given in Eq.1 and Eq.2. Although the former method has lower space complexity, since it only needs to interact with a spot of local neighbors, it seems not helpful to cut down the time complexity at all.

In summary, the AOC-YL method separate the communities more accurately but is inferior to the Affinity-propagation method in running speed. In the future studies, works still need to reduce the time complexity of the the AOC-YL method. Besides, improving the Affinity-propagation method to determine the number of exemplars more exactly is valued to be study in future as well.

6 Conclusions

In this paper, we give an review of some community mining methods in complex networks, including both the traditional methods and a series of new methods. In particular, we focus on introducing two Autonomy-Oriented Computing based methods referring to Affinity-propagation method and AOC-YL method. Finally, we realize the two methods on the same situations and got a comparison of their performances in two aspects: clustering quality and time complexity. We noticed that the AOC-YL method does better than the Affinity-propagation method, however runs much slower than the latter one. For future studies, reducing the time complexity of the AOC-YL method and improving the veracity of determine the number of exemplars of the Affinity-propagation method are both two valuable research directions.

References

- [1] Pothen A, Simon H, Liou K-P, "Partitioning sparse matrices with eigenvectors of graph", *SIAM J Matrix Anal Appl*, 11 (3): 430-452, 1990.
- [2] Guido Caldarelli, "Scale-Free Networks: Complex Webs in Nature and Technology", Oxford University Press, 38-57, May 2007.
- [3] Leon Danon, Jordi Duch, Alex Arenas and Albert Diaz-Guilera, "Community structure identification", 2005.
- [4] Radicchi F, Castellano C, Cecconi F, et al, "Defining and identifying communities in networks", *Proc Natl Acad Sci*, 101 (9): 2658-2663, 2004.
- [5] Gary William Flake, Steve Lawrence, C.Lee Giles and Frans M Coetzee, "Self-organization and identification of web communities", *Computer*, 35(3), 66-70, Mar 2002.
- [6] Brendan J. Frey and Delbert Dueck, "Clustering by passing messages between data points", *Science*, 315, 2007.
- [7] Palla. G, Derényi.I, Farkas I. and Vicsek.T, "Uncovering the overlapping community structure of complex networks in nature and society", *Nature* 435, 814-818, 2005.
- [8] M. Girvan and M. E. J.Newman, "Community structure in social and biological networks", *Proc. Natl. Acad. Sci. USA* 99, 7821-7826, 2002.
- [9] , Donetti L and Muñoz M. A, "Detecting network communities: a new systematic and efficient algorithm", *Journal of statistical mechanics: Theory and Experiment*, P10012, 2004.
- [10] M. E. J. Newman, "The structure and function of complex networks", *SIAM Review*, vol 45: 167, 2003.
- [11] M. E. J. Newman, "Detecting community structure in networks", *Eur. Phys. J. B* 38, 321-330, 2004.
- [12] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks", *Phys. Rev. E* 69, 026113, 2004.
- [13] M. E. J. Newman, "Fast algorithm for detecting community structure in networks", *Phys. Rev. E* 69, 066133, 2004.
- [14] J. R. Tyler, D. M. Wilkinson, and B. A. Huberman, Email as spectroscopy: Automated discovery of community structure within organizations. In M. Huysman, E. Wenger, and V. Wulf (eds.), *Proceedings of the First International Conference on Communities and Technologies*, Kluwer, Dordrecht, 2003.
- [15] Wasserman S and Faust K, "Social network analysis", Cambridge University Press, U.K., 1994.
- [16] BoYang, Jiming Liu, "An Efficient Probabilistic Approach to Network Community Mining", Book chapter, *Rough Sets and Knowledge Technology*, Springer, June 2007.

- [17] BoYang, William K. Cheung and Jiming Liu, “Community mining from signed social networks”, IEEE Transactions on Knowledge and Data Engineering, vol 19, Issue 10, 1333-1348, Oct 2007.
- [18] Jiming Liu, X. L. Jin, K. C. Tsui, “Autonomy Oriented Computing (AOC): Formulating computational systems with autonomous components”, IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans, vol, 35, no. 6, pp. 879-902, Nov. 2005.
- [19] Bo Yang, Jiming Liu, “An autonomy oriented approach to mining communities in distributed and dynamics networks”, IEEE International Conference on Self-Adaptive and Self-Organizing Systems (SASO'07), MIT 2007.
- [20] Xie Zhou and Wang xiaofan, “An Overview of Methods for Analyzing Community Structure in Complex Networks”, Complex systems and complexity science, vol 3, 2005.

Latent Topics Detection with Dirichlet Process Mixture Model

Zhan Tianjie

Abstract

In clustering, the exact number of cluster is always unknown. To deal with such problem, some nonparametric models are presented, such as Dirichlet process. Combined with mixture models, Dirichlet process could work better as it has a good inherent property of clustering. In this paper, Dirichlet process mixture model will be introduced, including the framework of the models, posterior distribution and inference.

A simple application with multinomial mixture is demonstrated later. The strong ability of clustering is shown by using the data from on line forum.

1. Introduction

In the field of machine learning, parameter selecting is one of the most important problem, especially the extract number of parameters. For direct using the known information to inference the unknown fact such as the incomplete data problem, nonparametric problem is introduced and the approach to that statistical problem, of which Dirichlet process is one successful model [1].

Dirichlet process can be traced to beta distribution and Dirichlet distribution with finite parameters, which has a property of exchangeability. According to Bayesian network, Dirichlet process can view as some kind of distribution over distribution. Similar to Gaussian process, which definite the distribution of functional space, Dirichlet process definite the distribution of some measure of distribution [4]. With some proper prior, dirichplet process can be used for Bayesian inference, parameters selection, model selection or density estimation and so on. This paper will focus on Bayesian inference, with a simple application for illustration.

Unlike traditional parametric model, Dirichlet process use unfixed and infinite number of parameters, with just a few in practical use, so that it can somehow over come the over- or under- fitting of data when there is misfit between the complexity of the model an the amount of data available. Traditional model use prior from a parameter family, which constraint distributions to lie within parametric families and limit the scope and type of inferences that can be made [2]. Alternatively, Dirichlet process inspired by nonparametric problem can cover the

whole space of distributions, so that inference of the posterior computations is tractable.

But unfortunately diriclet process mixture model can not use the EM method to make inference since it is infinite mixture models. So Gibbs sampling algorithm is used for inference based on the models, which could decrease the computations [6]. Similar to other conjugate priors such as beta distribution, Dirichlet process has a lower computational complexity and is popular in use in statistical learning field.

2. Dirichlet Process Mixture Model

Dirichlet Process (DP) as prior for some mixture models can be more flexible since it expands the space of available distribution space. Its most attractive point is that it works without knowing the exact number of cluster in advance so that it is well-worked when using proper models for mixing. Since DP can be inferences of finite mixture models which are simpler, following will start with finite case. A legible vision introduction of realization of DP called Stick-Breaking could refer to [1].

2.1 Definition

Formally, the Dirichlet process (DP) is a distribution over probability measure with probability one, which can be view as one kind of function space.

Given a dataset of $X := \{x_i, i = 1, 2, \dots, N\}$, the parameter θ and the conditional distribution of $p(x_i | \theta), x_i \in X$, we can assume the partitions of parameter θ follow the distribution G, which follow Dirichlet process, illustrated by figure 1.

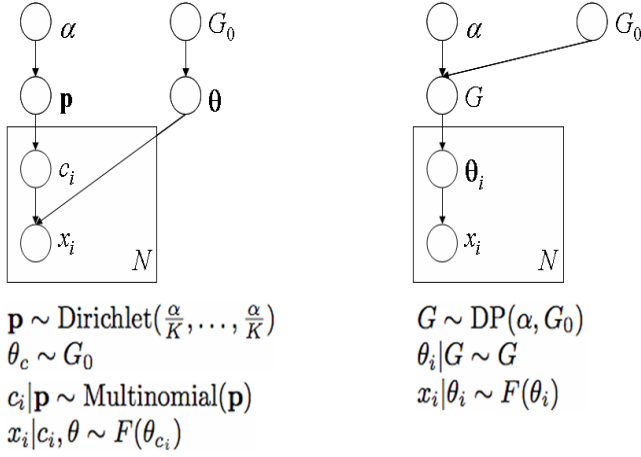


Figure 1 the Bayesian network box of Dirichlet process, the left one is the finite case, and the right one is the infinite case

So the Dirichlet process can be view as joining the mixing proportion parameter θ and the prior p with clustering variables c_i into a joint parameter θ_i^* .

Here the parameter α is a centralization parameter, which can adjust the chance that new cluster appears, and distribution can hold the information about clustering of the first (n-1) data, which can be used for make inference of the data x_i belonging to which cluster of known or a new cluster with the posterior distribution G . And G is assumed following DP. From the figure 1, it can be found that infinite case is accordant with the finite case, when the K is approaching to infinite. Since parameter θ denote the assignment to some cluster θ_k^* , so the scope of θ is $\theta^* := \{\theta_k^*, k = 1, 2, \dots, KK\}$, with KK represent the true number of cluster.

According to theory of DP, given $(\theta_1, \theta_2, \dots, \theta_n) \sim \text{DP}(\frac{\alpha}{K}, \frac{\alpha}{K}, \dots, \frac{\alpha}{K})$ with $K \rightarrow \infty$, x_i is¹ in the same cluster of x_j , that is $\theta_i = \theta_j$, with the mean probability of $\frac{\delta(\theta_j)}{N-1+\alpha}$ or in a new cluster with

the mean probability of $\frac{\alpha}{N-1+\alpha}$. So the θ_n can be estimated with the other known $\theta_j, j = 1, 2, \dots, n-1$ as

$$p(\theta_n | \theta_1, \theta_2, \dots, \theta_n) = \begin{cases} \frac{\delta(\theta_j)}{N-1+\alpha}, \exists \theta_j = \theta_i, j \neq i \\ \frac{\alpha}{N-1+\alpha}, \forall \theta_j \neq \theta_i, j \neq i \end{cases} \quad (2.1)$$

Here $\delta(\theta_i) = \sum_{j=1, j \neq i}^{n-1} 1(\theta_i = \theta_j)$, and $1(\theta_i = \theta_j)$

return 1 if true, 0 else.

Since the number of θ_i is infinite, (2.1) will be reformed as

$$p(\theta_i | \theta_{-i}) = \begin{cases} \frac{\delta(\theta_j)}{N-1+\alpha}, \exists \theta_j = \theta_i, \theta_j \in \theta_{-i} \\ \frac{\alpha}{N-1+\alpha}, \forall \theta_j \neq \theta_i, \theta_j \in \theta_{-i} \end{cases}, \quad (2.2)$$

here, $\theta_{-i} := \{\theta_j, j \neq i\}$

The process can be explained by Chinese restaurant algorithm.

2.2 Property of DP

2.2.1 Prior distribution

Let (Θ, B) be a measurable space, G_0 be a probability measure on that space, and α be a positive real number, with parameter $\theta_i \in \Theta$ So a Dirichlet process is any distribution of a random probability measure G over (Θ, B) such that, for all finite partitions $(\theta_1^*, \theta_2^*, \dots, \theta_n^*)$ of Θ ,

$$(G(\theta_1^*), G(\theta_2^*), \dots, G(\theta_n^*)) \sim \text{Dirichlet}(\alpha G_0(\theta_1^*), \alpha G_0(\theta_2^*), \dots, \alpha G_0(\theta_n^*)) \quad (2.3)$$

Here θ_i^* can be view as a cluster with

$$\theta_{i1} = \theta_{i2} = \dots = \theta_{in_i} = \theta_i^*$$

and $\delta(\theta_{i1}) = \delta(\theta_{i2}) = \dots = \delta(\theta_{in_i}) = n_i$,

Further, the parameters are exchangeable:

¹ The first θ is used in finite mixture model, and the second one in the infinite case.

$P(\theta_1, \theta_2, \dots, \theta_n) = P(\theta_{\pi(1)}, \theta_{\pi(2)}, \dots, \theta_{\pi(n)})$,
 here $(\pi(1), \pi(2), \dots, \pi(n))$ is any one permutation of $(1, 2, \dots, n)$

2.2.2 Posterior distribution

Because of the infinite exchangeability of θ_i^* , distribution (2.3) is the initial distribution, and the posterior distribution conditional on $(\theta_1, \theta_2, \dots, \theta_n)$ can be updated as when draw G from DP [4]:

$$p((G(\theta_1^*), G(\theta_2^*), \dots, G(\theta_n^*)) | \theta_1, \theta_2, \dots, \theta_n) \sim$$

$$Dirichlet(\alpha G_0(\theta_1^*) + \frac{\sum_{i=1}^n 1(\theta_i = \theta_1^*)}{\alpha + n}, \alpha G_0(\theta_2^*)$$

$$+ \frac{\sum_{i=1}^n 1(\theta_i = \theta_2^*)}{\alpha + n}, \dots, \alpha G_0(\theta_n^*) + \frac{\sum_{i=1}^n 1(\theta_i = \theta_n^*)}{\alpha + n})$$

So the posterior distribution of G conditional on $(\theta_1, \theta_2, \dots, \theta_n)$ is:

$$p(G | \theta_1, \theta_2, \dots, \theta_n) \sim$$

$$DP(\alpha + n, \frac{\alpha G_0}{\alpha + n} + \frac{1}{\alpha + n} \sum_{i=1}^n \delta(\theta_i))$$

and $E(G) = \frac{\alpha}{\alpha + n} G_0 + \frac{1}{\alpha + n} \sum_{i=1}^n \delta(\theta_i)$, here $\delta(\theta_i)$ is the 0-1 Dirichlet function.

2.3 Inference of DP

Generally, in the task of clustering, the assignment of data to proper cluster is the goal, especially the finding of hidden clusters. Since DP mixture models need not exact number of clusters, so the inference of number and the proportional value of component of mixture models can be done with well-built DP. However, the usual way of inference such as EM can not be used here, as DP is a nonparametric prior, and Gibbs sampling is used instead.

Cluster reassigning Algorithm with Gibbs sampling method is shown as:

(1) Reintroduce a cluster variable θ_i which takes on values that are the names θ_c of the clusters

(2) Store the parameters that are shared by all data in class θ_c in a new variable θ_c^*

(3) For $i = 1, 2, \dots, N$ sample x_i from

$$P(\theta_i = \theta_c | \theta_{-i}) = \begin{cases} \frac{\sum_{j \neq i} 1(\theta_j = \theta_c)}{N-1+\alpha} \int F(x_i | \theta) dH_{-i,c}(\theta), \exists \theta_j = \theta_c, j \neq i \\ \frac{\alpha}{N-1+\alpha} \int F(x_i | \theta) dG_0(\theta), otherwise \end{cases}$$

where $H_{-i,c}(\theta)$ is the posterior distribution of θ_c based on the prior G_0 and all observations for which $j \neq i$ and $\theta_j = \theta_c$

(4) Repeat

Since each of the parameter θ_i can viewed as the last one to be observed, and the N^{th} θ_i be estimated from posterior distribution conditional on other $N-1$ observant. However the Dirichlet distribution is discrete, so a kernel is needed to smooth out from draws from the DP to get a density distribution. In the case, the $F(x_i | \theta)$ plays as kernels indexed by θ [2].

According to the posterior inferred above, the estimated distribution is got with choice of proper base conditional distribution of x_i on parameter θ_i .

3. A simple application

Combine Dirichlet process mixture model of multinomial with prior of multinomial could get some good result in clustering of threads in an on-line forum. So following is a simple application of Dirichlet process mixture model with such a simple notion: assume

- (1) Threads in the on line forum cluster in the term of discussion on the same topics,
- (2) The latent topics among the threads following the distribution of multinomial,
- (3) The parameters θ_i in the definition of Dirichlet process denote which cluster belongs to.

3.1 Background and motivation

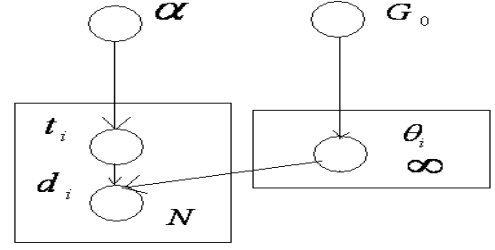
Since information on the web increase amazingly

and discussion on line is popular, most users spends lots of time in find few of information or interesting topics by reviewing large number of pages of posts. It is so time-consuming, and it is difficult to make a efficient search for some wanted topics sometimes. But traditional methods of clustering are helpless, since the numbers of pages is huge, and nobody knows how many topics indeed exist among the threads in one forum. So Dirichlet process mixture model is fit for that kind of case, as it does not need a known number of clusters before processing.

3.2 Proposed model

In the model, components for mixing are multinomial distribution, and prior is multinomial distribution, too. Since our dataset is use the frequency of participation of users over a thread, the real case is so similar to the Dirichlet process, which could be shown by the Chinese restaurant algorithm (CRA) [1]. In CRA, peoples always go the table with more people(samples), so the more crowded table will get more people to join, and the less on will get less and less. Similarly, users on the forum prefer to join the hit topics to the cooling. So in essence, Dirichlet process is matching the case on the forum well. Denoting d_i as the i_{th} threads, t_i as cluster parameters denoting which cluster belong to, θ_i as probability of one topic, which could explained as the degree of popular, and $(G_0(\theta_1), G_0(\theta_2), \dots, G_0(\theta_K)) := Multinomial(N; \theta_1, \theta_2, \dots, \theta_K)$.

Here K denote the expected number of cluster and N denote the number of data in the datasets. And the distribution of parameters θ_i , which follows $DP(\alpha, Multinomial)$, and the Bayesian graphics is shown in figure 2.



$$(G_0(\theta_1), G_0(\theta_2), \dots, G_0(\theta_K)) := Multinomial(N; \theta_1, \theta_2, \dots, \theta_K)$$

$$\theta \sim G_0, t_i | \theta \sim \theta$$

$$\theta | t_1, t_2, \dots, t_n \sim DP(\alpha + n, \frac{\alpha G_0 + \sum_{i=1}^n \delta(t_i = \theta)}{\alpha + n})$$

$$d_i | t_i, \theta_i \sim (Multinomial(\theta_i), t_i)$$

Figure 2 Bayesian network graphic of the model

3.3 Algorithms

From the model above, the main step for processing is as follow:

- 1) Initialize the DP mixture models with the prior G_0 , including the expected number of cluster K , the centralization parameter α
- 2) Initialize t_i with random number from 1 to K , and then initialize the value of θ_i with the initialized t_1, t_2, \dots, t_N
- 3) Repeat the Gibbs sampling to reassign the threads to clusters updated by number of times specified by user by Gibbs sampling:

Step 1 Randomly select one datum d_i and make sure each is selected only one time

Step 2 Delete the datum d_i from the former cluster, update the t_1, t_2, \dots, t_n , so t_i is unknown, and then update the $\{\theta_i\}$

Step 3 Calculate posterior probability of d_i conditional over the new $\{\theta_j\}, \{t_i\}$ with the density $P(d_i | \theta)$

Step 4 Use Gibbs sampling to get a sample t^i from the $P(d_i | \theta)$, and set $t_i = \lceil t^i \rceil$, and add d_i into the cluster of new t_i

Step 5 Update the $\{\theta_i\}$ with the new $\{t_i\}$

Step 6 Repeat until all the datum has been selected for only one time.

4) The clustering result could be given by clustering the data with the same t_i .

Here select the updating datum randomly could decrease the order of training data given.

4. Experiment of proposed algorithm

From the analysis of proposed model, it could be found the contribution of the algorithm and matching the data from real world well. So the data set is built with the data grabbed from the famous website of www.discussion.com.hk in Hong Kong, which is one of the main discussion forums in Hong Kong. And DP mixture model is implemented based on the Matlab package built by Yee Whye Teh [2].

4.1 Data set

In the experiment, user's participation model is used. One Thread is represented by a vector of frequency of participants in the thread. For filtering out of the noise produced by who always give large number of posts useless, threads is rebuilt with the famous text preprocessing method TFIDF. The centralization parameter α is set to be 1000, and expected number of cluster 50, repeat times of Gibbs sampling 60.

Totally, 2035 threads and 4483 users form the matrix of participation models, and each thread is represented by the participation of the 4483 users.

4.2 Some result

Here are 3 clusters randomly got from the resulting 13 clusters with one of that including few threads. Subjects randomly selected threads in the 3 clusters is shown in figure 3.

It could be found that in the first cluster "vehicles and the building" is discussed about, and "erotic" in the second cluster, "something unusual" in the third cluster.

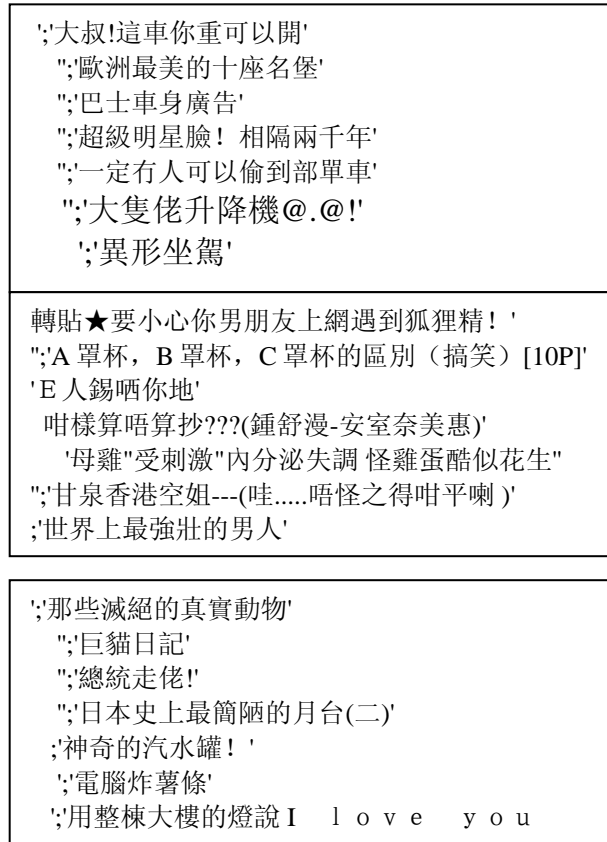


Figure 3 subjects of threads randomly selected from 3 clusters shown in each frame and each frame represents one cluster

Cluster	Numbers of threads
1	141
2	244
3	142
4	85
5	108
6	151
7	34
8	367
9	213
10	219
11	135
12	4
13	189

Figure 4 Totally 13 clusters are found and the distribution of threads among the 13 clusters are shown

4.3 Discussions

It could be seen in the frame above, threads can be clustered into different topics, since each topic is always associated with a group of users who share the common interest, with not true overlap between two groups if the topic is general enough.

The centralization parameter α could control the probability of new cluster, the larger, the more possible, and the smaller, the less possible. Contracted to the case without random selected threads to reassign, the result shows the result of clustering is less dependent on the giving order of data, and increasing the chance that threads of large cluster is updated so late that the cluster is divided into some other small cluster detected earlier.

5. Conclusion

Dirichlet process mixture model show a good ability of clustering especially in the case of unknown the exact number of clusters. And it is easy to compute the posterior distribution since it origin from beta distribution. Further, it is easy to understand using the model of CRA, and also build other models based on it quickly and efficiently.

6. Future work

Since the Gibbs sampling method used in the model is with data one by one, which is not only increase the complexity and also expand the influence of one data, and worse case when updating with some bad sample. So how to improve the sampling method is one way to improve the algorithm.

And further the distribution of data over the cluster parameters could be selected on a more space. By introducing fit functional space, it is expected that the prediction of cluster would be more exact, and increase the speed of approximation of the mixture model.

Some improved Dirichlet process could be used for improvement such as HDP, which build a larger distribution space and improve the flexibility of selection of distribution.

7. References

- [1] Teg Grenager, "Chinese Restaurants and Stick-Breaking: An Introduction to the Dirichlet Process", NLP Group Lunch, February 24, 2005.
- [2] Yee Whye Teh, "Dirichlet Process", Machine Learning Summer School 2007 Tutorial and Practical Course
- [3] T. S. Ferguson. "A Bayesian analysis of some nonparametric problems". *Annals of Statistics*, 1(2)209-230, 1973.
- [4] Aaron A. D'souza, "Notes on Dirichlet Process", http://www-clmc.usc.edu/~cs599_ct.
- [5] R. M. Neal, "Markov chain sampling methods for Dirichlet process mixture models". *Journal of Computational and Graphical Statistics*, 9:249-265, 2000
- [6] H. Ishwaran and L. F. James, "Gibbs Sampling methods for stick-breaking priors". *Journal of the American Statistical Association*, 96(453):161-173, 2001.

A Concurrent G-Negotiation Mechanism for Grid Resource Co-allocation

Benyun Shi

Abstract

Since computationally intensive applications may often require more resources than a single computing machine can provide in one administrative domain, bolstering resource co-allocation is essential for realizing the Grid vision. Given that resource providers and consumers may have different requirements and performance goals, successfully obtaining commitments through concurrent negotiations with multiple resource providers to simultaneously access several resources is a very challenging task for consumers. The novel contribution of this work is devising a concurrent mechanism that (i) coordinates multiple one-to-many concurrent negotiations between a consumer and multiple resource providers, and (ii) manages (de-)commitment for consumers during the one-to-many negotiation in Grid co-allocation. In this work, a utility-oriented coordination strategy and three classes of commitment strategies for concurrent negotiation are presented. A series of simulations were carried out in a variety of settings and favorable results show that the strategies (both commitment strategies and the coordination strategy) outperformed existing models in terms of utility, negotiation speed, and success rate.

1. Introduction

A Grid resource management system should bolster co-allocation of computing resources (i.e., allocating to an application multiple resources belonging to possibly different administrative domains) [5, p.2]. Supporting resource co-allocation is essential for realizing the Grid vision because (i) computationally intensive applications may require more resources than a single computing machine can provide in one administrative domain [4, p.1], and (ii) an application may require several types of computing capabilities from resource providers in other administrative domains. Sim [8] argued that software agents, in particular e-negotiation agents, can play an essential role in realizing the Grid vision. While there are existing works on applying bargaining to Grid resources allocation, very few works (e.g., [4][6]) considered negotiation for Grid resource co-allocation. Whereas SNAP (Service Negotiation and Acquisition Protocol) [6] searches for the solutions for satisfying simultaneous

multiple resources requirements of Grid consumers, it did not specify the strategies for negotiation agents.

Given that resource providers and consumers may have different requirements and performance goals, successfully obtaining commitments through concurrent negotiations with multiple resource providers to simultaneously access several resources is a very challenging task for consumers. Since there may be multiple resource providers providing a specific kind of resource, a consumer may select a required resource by adopting a one-to-many negotiation model. Additionally, for Grid resource co-allocation, resource selection would also involve coordinating multiple one-to-many concurrent negotiations and ensuring that the consumer can successfully obtain all required resources simultaneously. The impetus of this work is devising a concurrent negotiation mechanism that (i) coordinates multiple one-to-many concurrent negotiations between a consumer and multiple resource providers, and (ii) manages (de-)commitment for consumer in each one-to-many negotiation [1] [7] in which both consumers and providers can renege on a deal.

2. A Concurrent G-Negotiation Mechanism

This section describes an approach for the Grid resource co-allocation problem under a commitment model [1] [7] where renegeing on a deal is allowed for both consumer and provider agents. In this work the Grid resource co-allocation problem for n kinds of resources is transformed into a problem of n concurrent one-to-many negotiations where each one-to-many negotiation is also a concurrent negotiation for a particular kind of resource R_i , $1 \leq i \leq n$. Using this mechanism, a consumer in the Grid market negotiates simultaneously with multiple providers that supply possibly different types of resources. Denote $\{\mathcal{O}_j | 1 \leq j \leq n_i\}$ the set of n_i resource providers of the resource R_i , $0 < i < n+1$. Each consumer has n resources to acquire and a hard deadline τ_c for acquiring all n resources. Both an agent's preference for a resource and the strategy that it adopts during the negotiation are private information.

The negotiation mechanism consists of three components: a *coordinator module*, n commitment managers (CM_i , $0 < i < n+1$) and each CM_i manages a

number of negotiation threads. For each one-to-many negotiation for a particular resource, there exists a commitment manager agent CM_i [1] [7] that manages both commitments and de-commitments. Each CM_i adopts the management strategy to decide (i) whether or not to accept a resource provider's proposal or (ii) when to renege on a commitment at each negotiation round. Each negotiation thread (for a particular resource) follows a *Sequential Alternating Protocol*.

In this work, a consumer adopts three classes of time-dependent negotiation strategies (*Linear*, *Conciliatory*, and *Conservative*) in [3][8] to generate its proposals. Thus, three classes of commitment management strategies (CMS): {*Linear-CMS*, *Conciliatory-CMS*, and *Conservative-CMS*} can be specified and they correspond respectively to the *Linear*, *Conciliatory* and *Conservative* time-dependent strategies.

In [1][7], *degree of acceptance* is used to determine whether to accept a new proposal when there is no commitment for the consumer. Whereas [1][7] only consider the maximum predicted utility of the next proposal from other resource providers when they calculate the degree of acceptance, in this work, all proposals from resource providers at the current round are considered.

3. The Coordinator

The coordinator is used to determine when to terminate all one-to-many negotiation processes based on the information obtained from each commitment manager component so that the consumer's requirements and/or performance goals could be satisfied. In the Grid resource co-allocation problem, three factors are essential for a consumer: (i) successfully obtaining all required resources, (ii) obtaining the cheapest possible resource options, and (iii) obtaining the required resources rapidly. Since the failure of a one-to-many negotiation for any particular resource will result in the failure of the co-allocation for the consumer, ensuring a high negotiation success rate is the most important. In this section, a novel coordination strategy, called *utility-oriented coordination (UOC) strategy* is introduced to coordinate concurrent multiple one-to-many negotiations. In the *UOC* strategy, agents always prefer higher utility when they can guarantee a very high success rate.

During the co-allocation, once a resource provider's proposal is acceptable for a consumer (the proposal falls into the *agreement zone* of the consumer, i.e., $[IP_B, RP_B]$), it will be placed into an acceptable list for that resource by the consumer. If any acceptable list is empty, the coordinator cannot complete the co-allocation; otherwise, the coordinator can terminate all one-to-many negotiations based on its prediction of its utility of the coming round. At any round t , if there is no intermediate

deal in the sub-negotiation for resource R_i , the commitment manager in this sub-negotiation will predict all resource providers' possible proposals in the next round, and then calculate the expected gain ΔU_i^i by taking the difference between the maximal predicted utility of the next round $t+1$ (i.e., $\max_j \{U_{\text{exp}}^i(P_j^i(t)) | 1 \leq j \leq n_i\}$) and the maximal utility in the acceptable list (i.e., $\max_j \{U^i(P_j^i(t)) | 1 \leq j \leq n_i\}$). Hence,

the expected gain will be calculated as follows:

$$\Delta U_i^i = \max_j \{U_{\text{exp}}^i(P_j^i(t)) | 1 \leq j \leq n_i\} - \max_j \{U^i(P_j^i(t)) | 1 \leq j \leq n_i\}$$

where

$$U_{\text{exp}}^i(P_j^i(t)) = U^i(P_j^i(t)) + \frac{U^i(P_j^i(t)) - U^i(P_j^i(t-1))}{U^i(P_j^i(t-1)) - U^i(P_j^i(t-2))} \cdot |U^i(P_j^i(t)) - U^i(P_j^i(t-1))|$$

is the expected utility of the next proposal from resource provider O_j^i . Otherwise, if an intermediate deal has been established between the consumer and the owner O_k^i in the sub-negotiation at round t_{ik} , then at current round t , the commitment manager calculates ΔU_i^i by the possible utility loss at the following round:

$$\Delta U_i^i = U^i(\text{Avg}(P^i(t))) - U^i(P_k^i(t_{ik})).$$

All these estimated gains or losses will be submitted to the coordinator component. The coordinator then calculates $\Delta U_i = \sum_{i=1}^n \omega_i \Delta U_i^i$ where ω_i is the subject weight

of resource R_i of the consumer (the value of the weight may be affected by many factors, such as the market environment, the eagerness of the consumer to get the resource and so on, this will be studied in our future work). If $\Delta U_i < 0$, it means that the consumer will possibly lose some utility in the coming round(s), then the coordinator informs the commitment manager of each sub-negotiation that has not yet reached an intermediate deal to accept the best proposal from its acceptable list, and then terminate the entire negotiation.

4. Simulations and Experimental Results

To evaluate the effectiveness of the commitment management strategies and the coordination strategy of the concurrent negotiation mechanism in section 2, a series of experiments were carried out.

1) *Objectives and Motivations*: The objective of these experiments is to evaluate the performance of *Linear-CMS*, *Conciliatory-CMS*, and *Conservative-CMS* using the utility-oriented coordination strategy to coordinate concurrent negotiations under balanced Grid market situation. Furthermore, experiments to compare the *UOC*

strategy with the *patient coordination strategy* in [2] were also carried out.

2) *Experimental Settings*: The variables of these experiments are set as follows.

a) *Initial price and reserve price*: The initial price and reserve price are set to guarantee the existence of intersections between agreement zones (the domain between initial price and reserve price) of the consumer and that of each resource provider.

b) *Deadline*: The deadlines for each resource provider and consumer are uniformly generated from the time region [30, 80].

c) *Negotiation strategy*: All provider agents in this experiment make their proposals using time-dependant strategies [3]. Different providers have different time preference λ which is chosen from [0.1,10].

d) *Market type*: To simulate the complex real Grid environment, for each resource R_i , the number of resource provider (N_{RO}^i) is generated from a region [3, 10] (this is restricted by the computational capacity of the computer used for the experiments). The number of consumer is generated from the region $[2N_{RO}^i/3, 3N_{RO}^i/2]$ such that the ratio between the number of consumers and the number of resource providers is in the region $[2/3, 3/2]$. This kind of market is defined here as “balanced market”. Otherwise, if the ratio of number of consumers and the number of resource providers is less than 2/3 or greater than 2, the market is called “favorable market” or “unfavorable market”. (For instance, in an unfavorable market, there are too many consumers requiring for several resources, each of which only has few providers; and vice versa). Resource providers in this set of experiments renege from a deal based on the proposals and commitment time-periods of all consumers. Space limitation precludes details of renegeing on a deal from being described here.

3) *Performance Measure*: In the experiments, (i) the utility of the final co-allocation results, (ii) negotiation speed and (iii) success rate of acquiring all required resources are used as performance measures, defined as follows:

a) The utility of the consumer (U_c) is calculated by the following formula in the experiment:

$$U_c = \begin{cases} \frac{1}{N} \sum_{i=1}^N (U_c^i - \Gamma^i), & \text{if getting all resources} \\ 0, & \text{otherwise} \end{cases}$$

where $U_c^i = \frac{RP_c^i - P^i}{RP_c^i - MIN_o^i}$ is the utility of the consumer from a deal, (in which MIN_o^i is the minimum reserve price of all resource providers of resource R_i , P^i is the price of the deal), and Γ^i is the total penalty that the consumer should pay for resource R_i ;

b) The *speed* of the negotiation (S_c) is calculated by $S_c = t_c / \tau_c$, where t_c is the completion time of the negotiation, and a consumer prefers smaller S_c ;

c) Each experiment consists of 1000 runs, and the *final utility* and *speed* are averaged. The *success rate* is defined as the ratio of the successful negotiations over the total 1000 runs.

4) *Results and Observations*: Empirical results are shown in Fig 3.1-3.2. Fig. 3.1 shows the performances of the entire concurrent negotiation mechanism (final utility, speed, and success rate respectively) with the *UOC* strategy and *Linear-CMS*, *Conciliatory-CMS*, and *Conservative-CMS* in the balanced market situation. It can be observed that the consumer using *Linear-CMS* obtained the best utility while the consumer using *Conciliatory-CMS* achieved fastest negotiation speed. Both *Linear-CMS* and *Conciliatory-CMS* can guarantee a high success rate. This seems plausible because when the consumer adopts *Conciliatory-CMS*, it will make more concessions at each round, thus, it will be much easier to reach an agreement. At the same time, because it made large amounts of concessions, it will lose some utility during negotiation. When the consumer adopts *Conservative-CMS*, it cannot guarantee a high success rate because it makes a very small amount of concession at each round, which will further results in a lower final utility.

Fig 3.2 compares the performances of the *UOC* strategy with the patient coordination strategy in [2] using the same time-dependent commitment management strategies. This work adopts the *Linear-CMS* strategy because it is evaluated to be the best strategy among the three classes (*Linear-CMS*, *Conciliatory-CMS*, and *Conservative-CMS*) under the balanced market situation by empirical simulations. It was observed that the *UOC* strategy achieved better performance (higher utility, faster speed, and higher success rate). More importantly, it can be observed that the *UOC* strategy is more stable, i.e., the consumer obtained (almost) similar utility, speed, and success rate using the *UOC* strategy for different number of required resources. However, using the patient coordination strategy, the performance deteriorated with the number of required resources.

5. Conclusion

The contribution of this work is devising a concurrent negotiation mechanism (section 2) together with three classes of commitment management strategies and a utility-oriented coordination strategy (section 2.1) for managing multiple concurrent negotiations. The novelty and significance of this work is that it is among one of the earliest works (to the best of the authors’ knowledge) that consider a concurrent negotiation mechanism for Grid

resource co-allocation. This work only presents the empirical results of the G-negotiation mechanism in balanced Grid markets. Even though experiments have been carried out in both favorable and unfavorable markets, space limitations precludes empirical results for favorable and unfavorable markets from being included here. These empirical results, which were intended to show that the *Conciliatory-CMS* (respectively, *Conservative-CMS*) performs better in an unfavorable (respectively, a favorable market), will be reported in future papers.

6. References

- [1] T.D. Nguyen and N.R. Jennings, "Managing commitments in multiple concurrent negotiations", *Int. Journal Electronic Commerce Research and Applications*, 4 (4), 2005, pp. 362-376.
- [2] I. Rahwan, et al, "Intelligent agents for automated one-to-many e-commerce negotiation", *Twenty-Fifth Australian Computer Science Conf.*, 4, 2002, pp. 197-204.
- [3] K.M. Sim, "Equilibria, Prudent Compromises, and the 'Waiting Game'", *IEEE Transactions on Systems, Man and Cybernetics, Part B*, Vol. 35, No.4, Aug. 2005, pp. 712-724.
- [4] K.M. Sim, "Relaxed-criteria G-negotiation for Grid Resource Co-allocation", *ACM SIGECOM: E-commerce Exchanges*, Vol. 6, No. 2, Jan. 2007, pp. 37-46.
- [5] A. Ali et al., "A Taxonomy and Survey of Grid Resource Planning and Reservation Systems for Grid Enabled Analysis Environment", *Proc. of the 2004 Int. Sym. on Distributed Comp. and Appl. to Business Eng. and Sci.*, 2004, pp. 1-8.
- [6] K. Czajkowski, et al, "SNAP: A Protocol for Negotiation of Service Level Agreements and Coordinated Resource Management in Distributed Systems", *Job Scheduling Strategies for Parallel Processing*, 2002, pp. 1-10.
- [7] T.D. Nguyen and N.R. Jennings, "Reasoning about commitments in multiple concurrent negotiations", In *Proc. 6th Int. Conf. on E-Commerce*, Delft, 2004, pp. 77-84.
- [8] K.M. Sim, *G-Commerce, Market-driven G-Negotiation Agents and Grid Resource Management*, *IEEE Transactions on Systems, Man and Cybernetics, Part B*, Vol. 36, No. 6, Dec. 2006, pp 1381-1394.

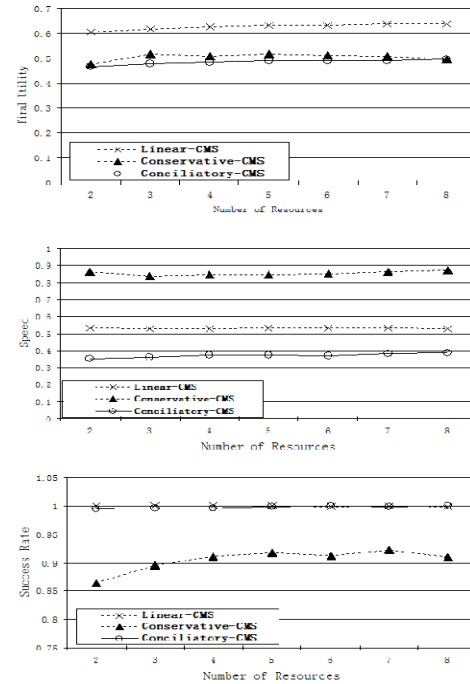


Fig. 3.1 Performances in Balanced Grid

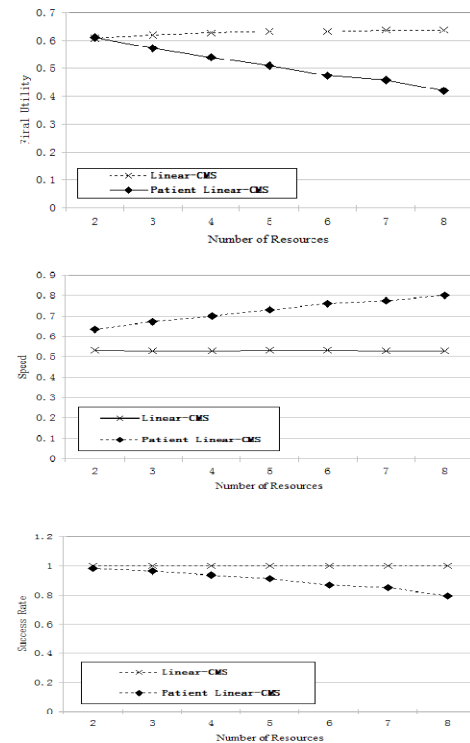


Fig. 3.2 Comparison of Coordination

Privacy Policy Enforcement in Service-Oriented Data Analysis Workflows

Kai-kin Chan

Abstract

There are several similarities between distributed data mining processes and scientific workflows. We model the distributed data mining processes as scientific workflows that can provide a better creation and management of data mining discovery pipeline. As the processes may involved different parties, privacy becomes an important issue. Policies we are used which can help the privacy enforcement. In this paper, we are going to demonstrate how to create ontologies for privacy aware data analysis, the methodology for creating domain specific data analysis workflow.

1 Introduction

Data analysis processes extract interesting patterns and discover useful knowledge from raw data. These processes may involve a number of data sets collected from different sources, and a number of analysis steps constructed by researchers of respective fields. For example, in medical research, patients' data could be distributed in different clinics and the medical researchers make use of various data analysis services provided by different research labs to complete some data analysis task. With the advent of the Internet, these "services" (or steps) can be invoked remotely via networked computer. While the execution of the data analysis processes can be managed by human, it's better to be managed with the help of the workflow management systems to achieve execution automation.

A workflow management system is a tool which enables users to compose and handle processes to achieve particular purposes, for example, data analysis. In general, workflows are composed of interrelated computational or data management jobs submitted to the remote hosts for execution to fulfill the goal of the workflow. Workflows have been used for modeling complex scientific applications. Such workflows are also called *scientific workflows*. With the recent advent of eScience, there is an increasing number of scientific applications that run on distributed computing environments. Modeling complex scientific applications as scientific workflows is an effective means for presenting and managing the execution of the underlying processes. In our

work, we model data analysis processes as workflows later on called data analysis workflows so as to take advantages of that.

The lifecycle of a workflow execution [2] includes steps to (1) create valid workflow description with respect to some domain independent and thus data independent constraints, (2) create workflow instances that specify the input data needed and the data flow among the analysis processes, (3) submit the fully described workflows for execution, and (4) schedule and dispatch jobs based on the dependency as specified in the workflows. It is easier to define, manage and execute the workflow followings these steps.

Data involved in data analysis processes could come from different parties and be sensitive as well, for example, where data privacy becomes an important issue. Policies described as constraints on the workflow can be used to control the access and usage of sensitive data. Our idea is to apply data privacy policies to data analysis workflows for managing data privacy. When the scale of data analysis becomes large, the effort needed manually for validating related workflows is tremendous. By adopting the semantic approach, data and analysis processes can be well described, searched and validated in a disciplined manner.

Wings [4, 3] is a workflow creation tool designed to create and validate very large scientific workflows. In Wings, workflow templates and instances are semantic objects. The workflow instances created by Wings can then be submitted to a grid-based distributed computing environment called Pegasus [1] for execution. Pegasus maps abstract workflows onto the grid environment. Physical locations for both workflow components and data are automatically located.

We are going to extend Wings that can support automatic privacy policy enforcement on data usage. By taking a semantic approach, privacy policies are represented using a semantic web language (OWL) and reasoned by a semantic web reasoner call Jena.

In our previous work, we extended the ontologies in Wings that can support the creation of data analysis workflows. In this paper, we are going to show how to create ontologies for privacy aware data analysis, the methodology for creating domain specific data analysis workflow.

2 Background

2.1 Wings and Pegasus

Integrated Wings [4, 3] into Pegasus [1] is an example of how semantic web technology can be used by workflow system. Workflow templates and instances are semantically described in Wings. The workflow instances can then be submitted to a grid-based distributed computing environment Pegasus for execution. Pegasus maps abstract workflows onto the grid environment.

Wings and Pegasus provide three stages in creation of workflows. The first stage is to compose workflow templates. Workflow templates specify the abstract structure of a workflow. It is a high level structure without identifying any particular data and resources. When composing a workflow template, user can access and search the existing workflow templates and component libraries in a particular domain. Experienced users can create and validate a workflow template. Less experienced users can search from the predefined workflow templates and specify the input data for execution.

The second stage is to create workflow instances. Workflow instances specify the data and resources used for the workflow template, such as the input data and output data. As the workflow template can be reused by different users, it is created as an abstract structure independent to the data and resources. Each time, users can process this stage to specify the resources to be associated with the template.

The third stage is to create an executable workflow. It involves the resources managements and data movements in the distributed services environment. The first and second stages can be done by Wings. After creating the workflow instances, Wings outputs a DAX (DAG XML description) file and a file library file. These files are XML based files which specify the inputs and outputs for the workflow, and then are submitted to Pegasus for execution.

In Wings, all the objects are processed as semantic objects with their metadata (properties) described in OWL-DL. They include components, files, collections, workflows and workflow templates. A set of ontologies representing these objects can be created for different domains. Then domain users can compose a workflow in Wings, and perform semantic checking on created workflow using Wings due to the ontology-based metadata.

3 Privacy Preservation and Data Analysis Ontologies for privacy policy enforcement

Wings is designed for creation and management of generic workflows. Fig. 1 shows how we extended the basic ontologies in Wings to include also concepts related to

privacy preservation and data analysis. The ontologies design is based on the basic ontologies in Wings. We have extended the basic ontologies by adding the most common data analysis concepts. The boxes in white background color represent the basic ontologies in Wings, the grey color boxes represent our extension of Wings ontologies.

Workflow Ontology (Extension): The following classes and properties are added.

- **Purpose** class and **for** property: It represents the data analysis purpose of the workflow template.
- **OutputQuality** class and **hasOutputQuality** property: It represents the overall output quality.

File Ontology (Extension): The following classes and properties are added.

- **DataSet** class: It extends the File class and represents the data sets.
- **ParameterFile** class: It extends the DataSet class and represents the parameter files for the processes.
- **Clusters** class: It extends the DataSet class and represents the intermediate forms of data products for later data analysis.
- **ClustersWithDataItems** class: It extends the Clusters class and represents the clustering result stored with data items.
- **ClustersWithStatistics** class: It extends the Clusters class and represents the clustering result stored with per-cluster statistics.
- **Attribute** class and **hasAttribute** property: It represents the attribute of data. not.

Component Library (Extension): It extends component ontology to represent data analysis processes. The following classes and properties are added.

- **Aggregate** class: It extends the ComponentType class. We add **GMMAggregate** and **DataSetAggregate** under Aggregate class.
- **DAComponentType** class: It extends the ComponentType class. We add **Clustering**, **ManifoldLearner**, **Classifier** and **AssociationRuleGenerator** under DAComponentType class. Each of these classes can have their subclasses such as **GMM-Basic**, **GMM-LFA**, **GTM**, **ISOMAP**, **SVM** and **C4.5**.
- **PPComponentType** class: It extends the ComponentType class. We add **Anonymizer**, **Generalizer**, **Encrypter** and **Perturbator** under PPComponentType class. Each of these classes can have their subclasses such as **k-anonymity**, **GMM-Abstract**, **addPerturbator**, **mulPerturbator**, **DES** and **RSA**.

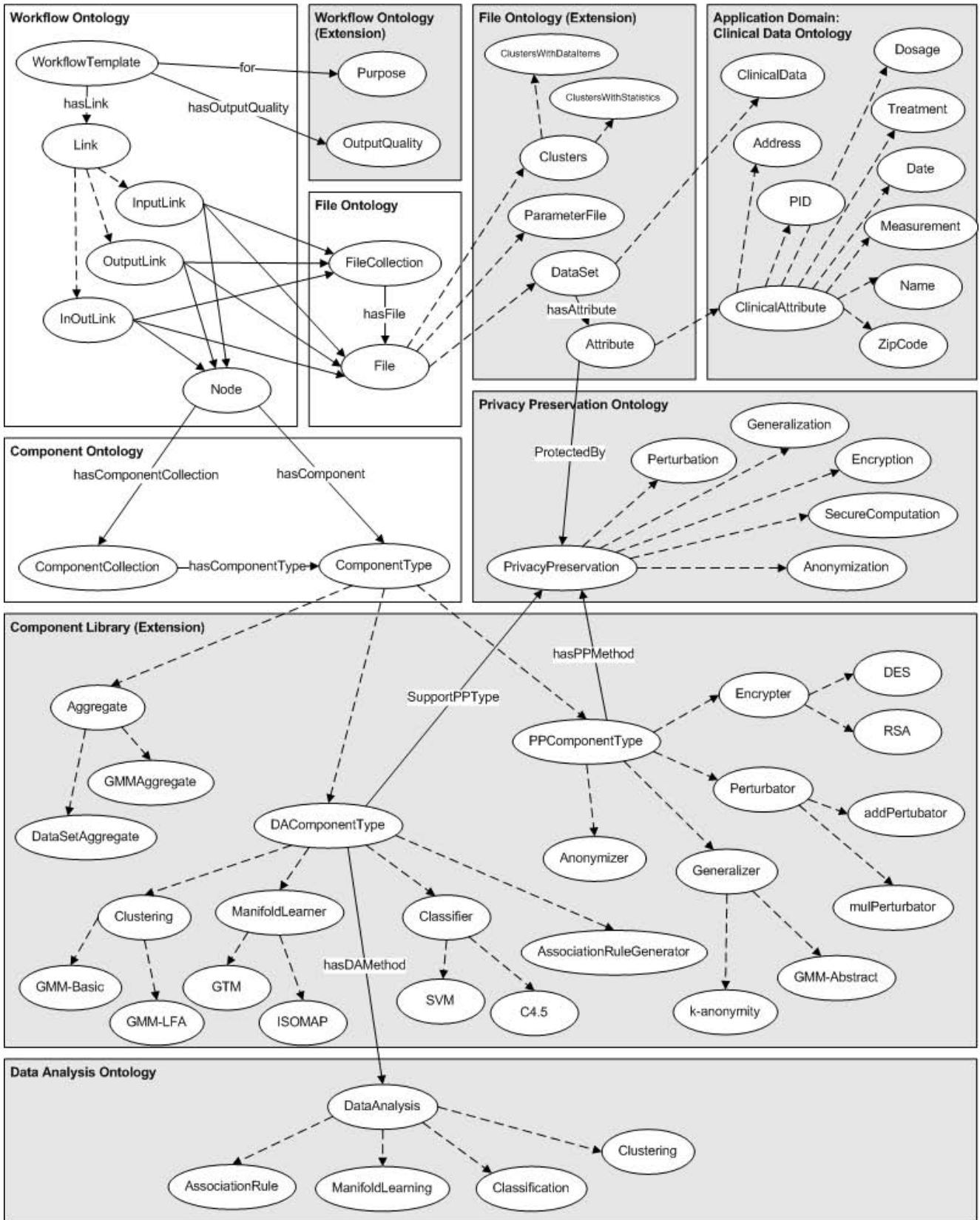


Figure 1. Ontologies for clinical data analysis

Privacy Preservation Ontology: The following classes and properties are added.

- **PrivacyPreservation** class: It represents the privacy preservation methods. Its subclasses contains **Perturbation, Generalization, Encryption, SecureComputation** and **Anonymization**.
- **SupportPPTType** property: It is to specify the privacy preservation method which a **DAComponentType** can allow its data to be processed.
- **hasPPMethod** property: It is to specify the privacy preservation method which a **PPComponentType** is implemented for.

Data Analysis Ontology: The following classes and properties are added.

- **DataAnalysis** class: It represents the data analysis methods. Its subclasses contains **AssociationRule, ManifoldLearning, Classification** and **Clustering**.
- **hasDAMethod** property: It is to specify the data analysis method which a **DAComponentType** is implemented for.

4 Methodology for creating domain specific data analysis workflow

The above extended ontologies are domain independent. They are the ontologies for privacy aware data analysis in general. It help the researchers to create an abstract structure of a data analysis workflow. This workflow does not specify the purposes, usages, data and resources, so we should create a domain specific workflow for policies checking. Fig. 2 shows the steps to create domain specific data analysis workflow. The first step is to create data ontology for application domain. The second step is to create workflow template. The third step is to create metadata and parameter files. The fourth step is to bind metadata and parameter files into workflow. The final step is to create and enforce policies.

In this section, we are going to show to create a domain specific workflow. We take clinical data analysis as an example.

4.1 Step 1: Creating clinical data ontology

As different application domain may contains its own attributes, so the clinical researcher may need to create a clinical data ontology which extends file ontology and specifies more detail about the clinical data. The following shows an example of clinical data ontology.

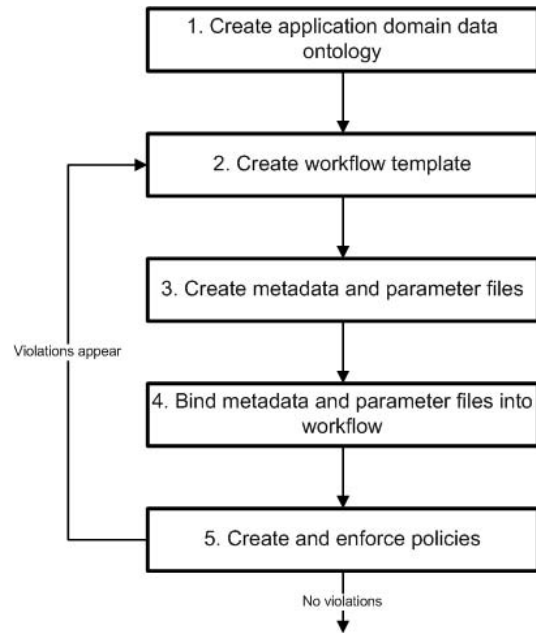


Figure 2. The steps to create domain specific data analysis workflow.

Clinical Data Ontology: The following classes are added.

- **ClinicalData** class: It extends the **DataSet** class and represents the data sets of clinical data.
- **ClinicalAttribute** class: It extends the **Attribute** class and represents the attributes of clinical data. It has some subclasses such as **Name, Address, Dosage**, they represent patients’ personal identification, demographic information and patients’ medical information.

4.2 Step 2: Creating workflow template

After creating clinical data ontology, the researcher need to create a domain independent workflow. Fig. 3 shows a domain independent data analysis workflow created by Wings. The workflow does not contain any information of clinical data, thus it is a generic workflow that does not specify the area of usage.

4.3 Step 3: Creating metadata and parameter files

As the data set input to the data analysis workflow are come from different clinics. Metadata are required to create from clinics for policies checking instead of sending to sensitive data set before execution.

A metadata of clinical data which must include:

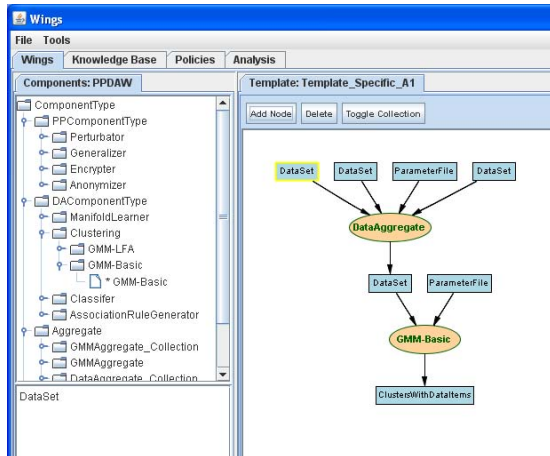


Figure 3. A domain independent data analysis workflow created by Wings.

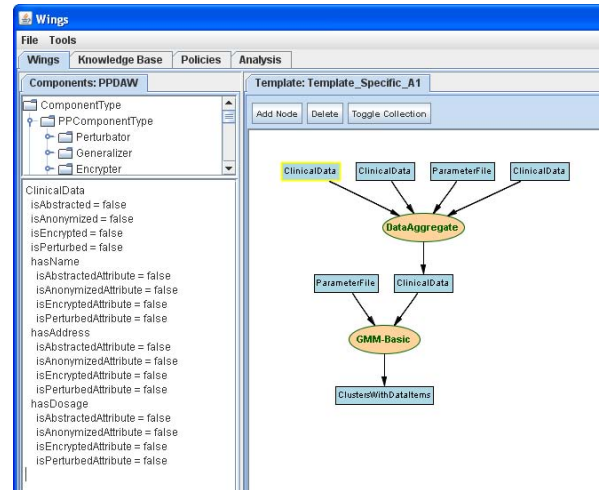


Figure 4. A clinical data analysis workflow created by Wings.

- the location of data sets.
- the specific domain of the data sets belong to and the attributes of data sets.

Moreover, the parameter files must be created by researchers that specifying which attributes should be proceed by components.

4.4 Step 4: Binding metadata and parameter files into workflow

The next step is to bind the metadata and parameter files into workflow. The binding procedure is to map the metadata and parameter files as inputs of workflow. Each data set would have 4 properties. They are (1) **isAbstracted**, (2) **isAnonymized**, (3) **isEncrypted** and (4) **isPerturbed** properties. And each data set would have several attributes which are specified in metadata. Each attribute set would also have 4 properties. They are (1) **isAbstractedAttribute**, (2) **isAnonymizedAttribute**, (3) **isEncryptedAttribute** and (4) **isPerturbedAttribute** properties. If all properties in a data set are set to false, it represents raw data. We assume all input data are raw data. Fig. 4 shows a clinical data analysis workflow created by Wings after metadata binding step.

4.5 Step 5: Creating and enforcing policies

There are some domain-specific policies for enforcing the domain-specific workflow.

Domain Specific policy S1: “For data that contain dosage information, it is not allowed that they are not first anonymized before being used for analysis.”
Context: $\text{hasLink}(?w, ?l) \wedge \text{hasFile}(?l, ?d) \wedge \text{hasAttribute}(?d, ?a) \wedge \text{Dosage}(?a) \wedge \text{hasDestinationNode}(?l, ?n) \wedge \text{hasComponent}(?n, ?c) \wedge \text{DACComponent}(?c)$
Usage: NULL
Protection: -ve: $\text{anonymized}(?d, ?aVal) \wedge \text{equal}(?aVal, \text{false})$
Correction: prompt [add an anonymization step right after (?d) found at (?l)]

Domain Specific policy S2: “For data that contain dosage information, it is required that they are first abstracted by k-anonymity before being further analysed.”
Context: $\text{hasLink}(?w, ?l) \wedge \text{hasFile}(?l, ?d) \wedge \text{hasAttribute}(?d, ?a) \wedge \text{Dosage}(?a) \wedge \text{hasDestinationNode}(?l, ?n) \wedge \text{hasComponent}(?n, ?c) \wedge \text{DACComponent}(?c)$
Usage: NULL
Protection: +ve: $\text{abstracted}(?d, ?dgVal) \wedge \text{equal}(?dgVal, \text{true}) \wedge \text{abstractedBy}(?d, ?m) \wedge \text{k-anonymity}(?m)$
Correction: prompt [add an k-anonymity step right after (?d) found at (?l)]

Fig. 5 shows the workflow checking results in Wings. The results would show what correct actions should be done to correct the workflow, so the analyst can go to modify the workflow and bind the metadata again, finally run the policies checking to ensure there are no violations.

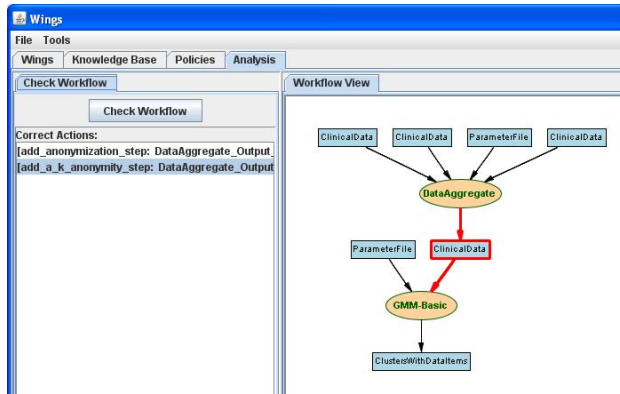


Figure 5. Workflow checking results in Wings.

5 Implementation

In the previous section, we have created the clinical data ontology in step 1. Fig. 3 shows that we are going to create a domain independent workflow template. Fig. 4 and Fig. 5 show that after binding and policies checking, there are 2 violations in the workflow.

When there are some violations, the researchers can go back to step 2 to modify the workflow template, and bind again the new metadata files. Fig. 6 shows the updated workflow.

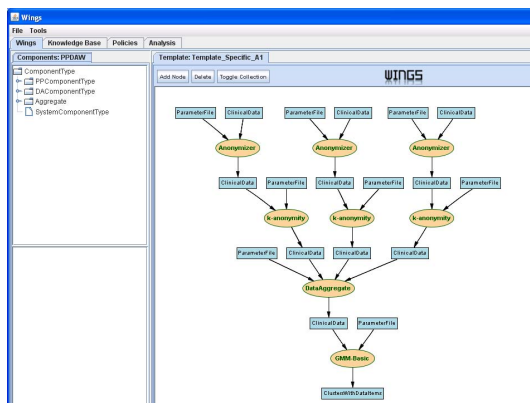


Figure 6. Updated Wings workflow.

6 Conclusion and Future works

In this paper, we create ontologies for privacy aware data analysis, design the methodology for creating domain specific data analysis workflow. We integrate the policy reasoning and checking part into Wings, and implement the enforcement part that can automatic protect the privacy data.

Our future work is to extend Wings that can support automatic correction to the workflow. Where there are some violations after policies checking, the researchers need to modify the workflow and do the policies checking again. In the future, automatic correction to the workflow will be implemented, so it is easy for researches to correct the policies violated workflow. We also motivated to extend Wings that can work well for BPEL4WS environment. Wings is original designed for Pegasus. But Pegasus is not designed under SOA. In some specific analysis, it is also common for researchers to add their own analysis processes. SOAs specified the standard input and output format of processes, it leads the system provide certain extensibility. Researchers could implement their own analysis processes that can add and execute in existing workflow with the help of SOAs. In the future, we will implement Wings for BPEL4WS which is an industry standard for describing web services-based business processes. We will also test the performance and scalability of the BPEL4WS workflow. We aims to shorten the execution time to increase the performance, and handle large amount of data.

References

- [1] Ewa Deelman, James Blythe, Yolanda Gil, Carl Kesselman, Gaurang Mehta, Sonal Patil, Mei-Hui Su, Karan Vahi and Miron Livny: Pegasus: Mapping Scientific Workflows onto the Grid, 2004.
- [2] Ewa Deelman, Yolanda Gil: Managing Large-Scale Scientific Workflows in Distributed Environments: Experiences and Challenges. Proceedings of the Workshop on Scientific Workflows and Business Workflow Standards in e-Science, The Second IEEE International Conference on e-Science and Grid Computing, 2006.
- [3] Jihie Kim, Yolanda Gil and Varun Ratnakar: Semantic Metadata Generation for Large Scientific Workflows. Proceeding of 5th International Semantic Web Conference, ISWC-2006, 2006.
- [4] Yolanda Gil, Varun Ratnakar, Ewa Deelman, Marc Spraragen and Jihie Kim: Wings for Pegasus: A Semantic Approach to Creating Very Large Scientific Workflows. Proceeding of OWL: Experiences and Directions, 2006.

Goal Oriented Requirements Engineering – Goal Definition

Wu Di

Abstract

The goal technique has been well developed up to now. However, there is not any specific method to indicate “when”, “where” and “how” the goal technique could be used. Therefore, this paper proposes goal-oriented system life cycle (GOSLC) architecture to move goal technique into practice. Based on the results of GOSLC architecture, requirements could be defined by using the definition of goals. Moreover, methods for goal definition have been illustrated. The methodology was evaluated via a case study of online banking system. The results show the effectiveness of the proposed methodology.

1. Introduction

When examining the maturity of goal technique, it is necessary to check if there is a ground to move goal technique into practice (Redwine, Riddle, 1985). Other researches for goal technique enhancement are mainly focuses on goal definition, goal evolution etc., but lacking of moving goal technique into practice. System define life cycle is widely used for system development. We combine goal technique with system development life cycle, by using goal definition to define system requirements.

In this paper, GOSLC architecture was proposed. The proposed methodology is a technological approach to defining and validating system requirements by using goals. This could not only used for technical system but for social system, as none specific work platform required. Also, in order to make our methodology non-professional-knowledge based, we elicit goals by role of people and reconcile goals by negotiations. The effectiveness of the methodology based on a case study of online banking system is shown.

The rest of this paper is organized as follows: Section 2 describes the background of goal technique and reviews several goal-oriented methodologies. Section 3 discusses the GOSLC architecture. Section 4 introduces our non-professional-knowledge based idea for goal definition. Section 5 proposes a case study for our proposed methodology. Section 6 makes a conclusion and talks about the future work

2. Background and Literature Review

In the early 1990s, the importance of the goal concept has been identified, e.g. (Yu, 1993). Myriads of system development methodologies using goal concept have been proposed. The goal concept among the proposed methodologies is not limited to the original definition of goal – desired outcome (WIKIPEDIA, 2007), but extended to several aspects. One is making clear the purpose of the system, e.g. in (Rolland et al., 1999), goal is defined as “something that some stakeholder hopes to achieve in the future”. Another is guaranteeing requirements quality, e.g. in (Anton, 1996), goal is taken as a tool to identify, organize, and justify software requirements. Also aiding correct implementation of system functions was noticed, e.g. in (Lamsweerde et al., 1993), achievement goal was proposed to formulate implementation.

Redwine and Riddle (Redwine, Riddle, 1985) indicate that a technology typically takes 15 to 20 years to be ready for popularization, and several typical phases – basic research, concept formulation, development and extension, internal enhancement and exploration, external enhancement and exploration, and popularization, could be followed. Here, we use this method to describe the popularization of goal. In the basic research period (1985-1993) of goal oriented requirements engineering (GORE), goal was only taken as a part of supporting components in the requirements definition. For example, in (Mylopolos et al., 1992), goal was used for identifying non-functional requirements, such as performance, operational costs, and constraints, in order to construct the framework for integrating non-functional requirements into the software development process. In the concept formulation (1992-1996) and development and extension (1995-2000) periods of GORE, goal was regarded as the main method for requirements definition, as in (Anton, 1996), requirements were identified by a process – identifying, organizing, refining, and operationalizing goal. The goal technique was well-organized in the late 1990s. Later researches focused more on the goal technique enhancement, making goal concept popular and applying goal concept as a mature technique for the requirements engineering. For example, in A Unifying Framework (Kavakli, 2002), a framework for requirements definition was proposed by accepting several goal-based methodologies, such as KAOS that identifies responsibilities and assigns operations of an information

system by elaborating goal structure. The goal structure includes the steps of identifying goals, which is giving names, and classifying goals, and structuring goals, which is refinements, and solving conflicts, (Lamsweerde et al., 1993).

The goal definition, which is an indispensable step in goal oriented requirements engineering (GOREs), directly affects or even determines the goal-based methodology. For example, in Goal Scenario coupling (Rolland et al., 1999), requirements could not be identified without goal identification and analysis. Also in (Briand, 2002), the author explicitly indicated that goal definition is the fundamental step in their process for defining measures, which are abstracted based on goals.

Many researches have been done on goal definition. In order to get the detailed description on goal definition, surveyed about goal definition on several typical goal-based methodologies as shown in Table 1. We organized them using the technology popularization theory (Redwine, Riddle, 1985) introduced earlier. In Table 2, we provide a brief overview of each surveyed methodology.

Goal definition methods in these surveyed methodologies are introduced in Table 3.

In Table 3, the process of most goal definition methods in surveyed methodologies is setting goals or defining goals, and then refining goals. We could abstract that mainly two steps are considered in goal definition process – goal elicitation and goal reconciliation. Explanations of the proposed two steps are shown as below:

- Goal elicitation – a process for identifying goals from the existing information, which can be accessed by developers.
- Goal reconciliation – refining elicited goals.

In Table 4 and Table 5, we illustrate the techniques in goal elicitation and goal reconciliation from the surveyed methodologies separately to make further introduction.

In Table 4, most surveyed methodologies prefer to elicit goals from several defined aspects, such as in GQM, goals are elicited from products, processes, and resources. Also decomposition is another method that frequently used. For example, GSN supposes some goals are existed. Based on these existed goals, sub-goals are elicited “until a point is reached where claims can be supported by direct reference to available evidence” (Kelly, 1998). For example, the one of the existed goal in online banking system is the online banking system should be safe. Sub-goals could be elicited as 1) The online banking system complies with relevant safety requirements. 2) The online banking system offers enhanced safety over existing systems. For sub-goal 2, it could be ended with argument over each safety features. Eliciting goals in hierarchy is used in Goal Scenario Coupling, in which goals are first

elicited from contextual level, then from functional level, at last from physical level.

From Table 5, we could learn that negotiation is a main method for goal reconciliation, such as in KAOS, in which goals are reconciled by negotiating conflicts among them. Constraint is another tool for goal reconciliation, such as in GBRAM. FBCM refines goals by facts, which are “observed results occurring in business fields” (Kokune et al., 2007).

Table 4 and Table 5 contain several goal definition methods. But these methods are isolated. By using these methods, we could define goals. However, when, where, and how are these defined goals used? We have illustrated earlier that the goal technique becomes mature. Now we make the enhancement of the goal technique. Redwine and Riddle (Redwine, Riddle, 1985) indicate that the responsibility of maturity is not just that new ideas are promising, but also that they are effective (a necessary ground to move into practice). To move goal definition method into practice, the supports for facilitating the following activities should be contained:

- Activities for applying goals to the phases of system life cycle.
- Activities for deriving system physical components by using goals.

After applying goals to the phases of system life cycle, we know when to use the defined goals. Then we identified that these defined goals should be used for deriving system physical components, which describe a system physically.

Later, an architecture to demonstrate the relationships among goal, system life cycle, and system physical components, will be proposed.

After we make sure the goal definition method in our research is practical, specific method for goal definition would be shown.

The surveyed methods on goal definition somewhat calls for professional knowledge on specific points, such as eliciting goal from the process aspect. Users have to know what process is. So a method that defines goals by different worldviews is useful. People who use this method are not required specific background.

3. Goal-oriented System Life Cycle Architecture

Our proposed architecture, which is designed to describe system life cycle by using goals, and derive system components from goals is proposed in Figure 1.

In system life cycle, strategic study, system planning, system analysis, system design, construction design, construction and workbench test, installation, test of installed system, operation, evolution, phase out, and

postmortem, are included. (Olle et al., 1991) Here, we mainly consider the system planning, system analysis,

Technology popularization stage	Basic research and concept formulation period			Development and extension period		Integration period	
Methodology	NFR (Mylopoulos et al., 1992)	KAOS (Lamsweerde et al., 1993)	GQM (Basili, 1994)	GBRAM (Anton, 1996)	GSN (Kelly, 1998)	Goal Scenario Coupling (GSC) (Rolland et al., 1999)	FBCM (Kokune et al., 2007)

Table 1 – Goal-based methodologies to be illustrated

Methodology	Overview
NFR	Using goals, link types, methods, correlation rules, and the labeling procedure to represent and use non-functional requirements.
KAOS	A multi-perspective language, as it fulfills the rich ontology. A goal driven elaboration method, as it composes four models, including goal.
GQM	They ask questions, give answers, and map with the real situation to improve or as a principle.
GBRAM	Discuss goals from goal analysis and goal evolution.
GSN	A graphical argumentation notation -explicitly represents the individual elements of any safety argument and the relationships that exist between these elements.
GSC	Use the CREWSL' Ecrire approach to couple goals and scenarios together.
FBCM	A technological approach to defining non-functional requirements that are used to set the business goals.

Table 2 – Overview of these goal-based methodologies

Methodology	Goal definition method
NFR	<ul style="list-style-type: none"> ● Set goals for representing NFR, design decisions and arguments in support of other goals. ● Provide link types for describing relating goals or goal relationships. ● Refine goals by Decomposition, Satisfying, and Argumentation methods.
KAOS	<ul style="list-style-type: none"> ● Identify goals by goal classification. ● Refine goals by decomposition.
GQM	Identify goals from: Purpose, issue, Object (process), Viewpoint.
GBRAM	<ul style="list-style-type: none"> ● Extract goals from various types, such as flow charts, ERD. ● Identify responsible agents of goals. ● Identify constraints by searching for temporal connectives to refine goals. ● Consider the goal precedence relations for refinement.
GSN	Decomposing goals until a point is reached where claims can be supported by direct reference to solutions.
Goal Scenario Coupling	<ul style="list-style-type: none"> ● Define goals from contextual, functional, and physical level. ● Use template to reformulate the previous informal goal into a more accurate definition. ● Scenario is authored after the goal is selected. ● Elicit goal through scenario analysis.
FBCM	<ul style="list-style-type: none"> ● Strategic goals are extracted from existing papers such as actions plans. ● Refine strategic goals by facts: additional goals are discovered by observing the actual field.

Table 3 – Introduction to the goal definition method in these goal-based methodologies

Methodology Method	NFR	KAOS	GQM	GBRAM	GSN	Goal Scenario Coupling	FBCM
Decomposition	X	X			X		X
Elicit goal in hierarchy						X	
Elicit goal from several defined aspects.		X	X	X		X	X

Table 4 – Goal elicitation methods used in these goal-based methodologies

Methodology Method	NFR	KAOS	GQM	GBRAM	GSN	Goal Scenario Coupling	FBCM
Negotiation	X	X			X		
Constraints				X		X	
Refine goals by facts							X

Table 5 – Goal reconciliation methods used in these goal-based methodologies

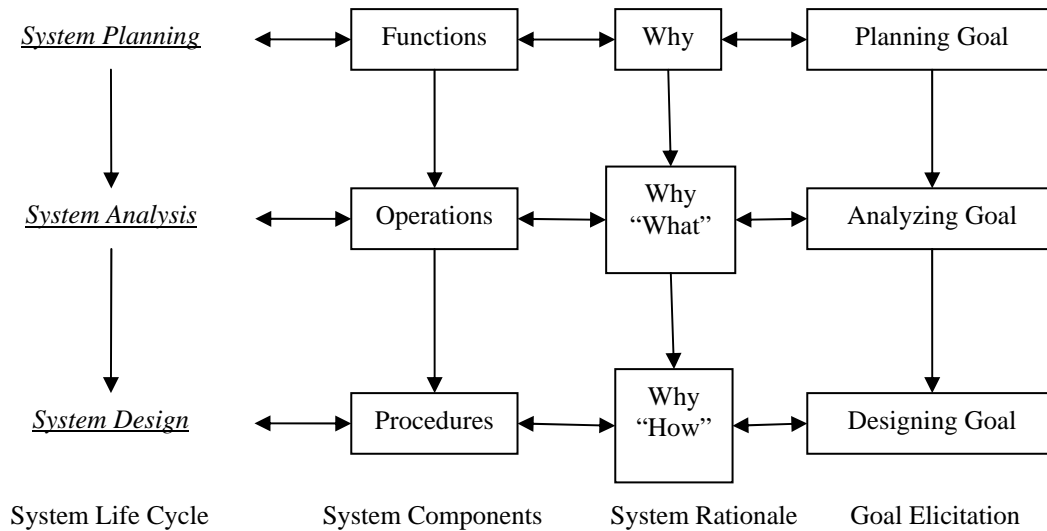


Figure 1 – Goal-in-System Architecture

System Life Cycle	System Components	System Rationale	Goal Definition
System planning	Prospected functions have been defined.	Why a system should be developed? Why certain function is needed?	Planning Goal: Making clear the purpose of a system
System analysis	Operations implement functions.	What processes should we follow to realize the function(s)?	Analyzing Goal: Guaranteeing requirements quality
System design	Procedures provide detailed information for implementation.	How to prepare for implementing the developing method?	Designing Goal: Aiding correct implementation of system function

Table 6 – Mapping Components in Goal-in-system Architecture

system design stages. In system planning, the broad nature of the information requirements in the enterprise is determined, business objectives have been identified. Also a feasibility study will be done to determine the possible alternatives for proceeding further. (Olle et al., 1991)

In system analysis, a problem solving technique that decomposes a system into its component pieces for the purpose of studying how well those component parts work and interact to accomplish their purpose. (Bentley and Whitten, 2007)

In system design, a complementary problem-solving technique that reassembles a system's component pieces back into a complete system. This may involve adding, deleting, and changing pieces relative to the original system. (Bentley and Whitten, 2007)

According to the three phases of system life cycle, we proposed three kinds of corresponding goal – planning goal, analyzing goal, and designing goal, to describe the responsibilities of the phases, which have been introduced before.

Based on the goals we have defined, system components will be derived to physically describe a system.

“System rationales can include not only the reasons behind a design decision but also the justification for it, the other alternatives considered, the tradeoffs evaluated, and the argumentation that led to the decision” (Lee, 1997). So we added the system rationale – why, why “what”, and why “how”, to improve dependency management, collaboration, reuse, maintenance, learning, and documentation (Lee, 1997).

System Life Cycle – System Planning-System Analysis-System Design

In system planning, the reason for developing prospective function(s) should be identified. A more in-depth explanation for changing existing system or developing a new system should be illustrated.

After we have fixed the changing component(s), in system analysis, specific methods for operating a system should be provided. Also we make clear the boundary of the changing component(s).

Finally, in system design, details should be abstracted for system construction from the information we have handled.

System Components – Functions-Operations-Procedures

If we view a system development physically, functions, which is defined as “an attribute of an object that describes what the object does, the phenomena it exhibits, the service it supports, or what it is used for (Coderborg, 2003)”, should be identified first, as we make clear the

purpose of developing a system. After Functions are realized, Operations are derived for implementing functions, such as to create a business need/objective (Chan, 2000). Procedures define detailed information as steps to follow to implement (delineating what have to be done) an operation.

System Rationale – Why-What-How

Some questions should be answered to illustrate the system rationale issue. Why should a system be developed? Why is a certain business function needed? Why should such kinds of processes be followed to realize the functions (business) (why what)? Why should we prepare for implementing the processes in some particular ways (why how)?

Goal Elicitation – Planning Goal-Analyzing Goal-Designing Goal

The three proposed goals can be defined as below:

- Planning Goal – Making clear the purpose of a system
- Analyzing Goal – Guaranteeing requirements quality
- Designing Goal – Aiding correct implementation of a system

Hierarchy existed in the three kinds of goal. Analyzing goal is elicited from planning goal, and designing goal is elicited from analyzing goal.

Mapping Components in GOSLC Architecture

In the system planning phase, planning goal is defined to make clear the purpose of a system. At the same time, the reasons why the prospective system should exist have been identified. Then system functions could be derived to show the responsibility of the system.

In the system analysis phase, after we have confirmed the functions and purpose of a system, analyzing goals should be proposed to guarantee the requirements quality. Also, we should identify what processes we should follow to realize the functions. The rationale issue considers about “what” should be shown. Finally, operations could be derived to implement functions.

In the system design phase, by using the fruits from system analysis, designing goals could be defined to aid correct implementation of system functions. The “How” issue will provide us with ideas for implementation. And the rationale issue based on “how” will make the implementation more feasible. Finally, procedures could be derived to provide detailed information for implementation.

4. Goal definition

As we have introduced before, two steps are included in goal definition – goal elicitation and goal reconciliation. We could not separate the two steps completely, as we could not elicit goal fully in a cycle. There are iterations.

For planning goal, we elicit them from role of people, which is defined by the method in soft system methodology (SSM). (Checkland, P, 1988, 1999) SSM is used for ill-defined system, with situations that are difficult to understand and to define. In SSM, different world views are constructed by using CATWOE (Checkland, P, 1988):

- Clients or Customers - Those who are benefited or are affected
- Actors - Those who carry out the system activities
- Transformation - Changes within or because of the system
- Worldview - How system is perceived, making T meaningful in context
- Owner-Those who could stop T
- Environment - The world that surrounds and influences the system, but the clients, actor and owner have no control over it. The environment is taken as given.

We take three kinds of role of people – customer, actor, and owner to elicit planning goal. Each kind of role of people makes clear their purpose towards the developing system.

After we have finished eliciting planning goal from role of people, we elicit planning goal from non-functional requirements (NFRs), such as security, performance, cost, and accuracy. Also NFRs could do the goal reconciliation. Specific illustration on NFRs will be introduced in future work.

As some people know why they propose a certain purpose of a system, but do not know how to implement it. Analyzing goals are proposed to guarantee requirements quality (achieve the planning goals). The way for checking the availability and efficiency of analyzing goal is that having the analyzing goal, the planning goal could be achieved.

Both planning goal and analyzing goal are used for defining system requirements. Then designing goal would be defined for aiding correct implementation of the developing system. The way for evaluating designing goal is the same as evaluating analyzing goal – having the designing goal, and then analyzing goal could be achieved.

But the levels for goal achievement are different. For example, the planning goal is the online banking system should be safe. The analyzing goal for this planning goal could be:

- 1) 1.1 login system is provide
- 2) 2.1 login system is provide

2.2 special plug-in components is set up for the login system

Obviously, difference existed between the two kinds of analyzing goals. Which should be adopted for the further explanation depends on other planning goals and non-functional requirements. This is the issue that we should solve in goal reconciliation, which will be worked out.

5. Case Study

We take an online banking system as the case for the feasibility demonstration of the proposed methodology. For short, each situation only has one example.

Elicit planning goal from existed information

1. Elicit planning goal from role of people:

Role of people		Planning goal
Clients	on-line banking users	1. View online account balance
Actors	administrator	2. Transactions are processing safely
Owner	bank	3. The on-line banking is helpful for the bank

Table 7 – Elicit planning goal from role of people

2. Elicit planning goal from Non-functional requirements:

Non-functional requirements	Planning goal
Security	4. The system is secure
Performance	5. The system is user-friendly
Cost	6. Low cost (time, people, and money)
Accuracy	7. The system is accurate
Reusability	8. The system will not be reused

Table 8 – Elicit planning goal from non-functional requirements

The planning goals elicited from non-functional requirements are mainly determined by owners, such as 6 and 8. But owners have rights to let other kinds of role of people join in the discussion. For example, the bank could collect the opinions from the ordinary users for building the online banking system.

Some planning goals elicited from non-functional requirements would be the same as elicited from role of people, such as 2 and 4. In this situation we will combine them as one planning goal.

Here, we could get the purpose of the system is the planning goal 1, 2, 3, 5, 6, 7, 8.

Derive system functions from planning goal

From planning goal 1, a function “show online account balance” is derived.

From planning goal 2, a function “security function” is derived.

The same way is used for deriving other functions. Also planning goal could be taken as a tool for checking “why” such function is provided by the developing system.

Elicit analyzing goal from planning goal

To achieve planning goal 1, we have the analyzing goal “1) show online account balance to the user”. If this analyzing goal is realized, the planning goal 1 will be achieved. Also these kinds of analyzing goal could be directly elicited from planning goal.

To guarantee the quality of planning goal 2, discussion should be done between developers and owners. Some other planning goals should be considered, as we could not fix the level of system security, which requires a negotiation with planning goals 6 and 7. Here, we provide two analyzing goals to achieve planning goal 2 – 2) have login system, and 3) set up plug-in components for the login system.

Derive operations from functions and analyzing goal

The function 1 can generate the operation “1) show online account balance”.

For the function 2, we do not know “what” kind of secure functions the bank needs. But analyzing goals can help. From analyzing goal 2 and 3, operations for implementing “secure function” can be derived – 2) users should login, and 3) users should set up plug-in components. Analyzing goals show “why” operations 2 and 3 (what) are derived.

Define designing goal from planning goal and analyzing goal

For the analyzing goal 1, we have the designing goal “1) show online account balance to the user”.

For the analyzing goal 2, we can not directly elicit designing goal to aid the implementation for users to login, as we could not fix the login steps. So discussion should be done here to figure out the steps. Also other analyzing goals should be considered, such as analyzing goals elicited from planning goal 6 and 7. Here we provide one kind of situation – 2) users type in user name, and 3) users type in password.

Derive procedures from operations and designing goal

The operation 1 could generate procedures 1) select “show the balance of account” 2) show account.

For the operation 2, we could not directly generate procedures. But designing goal could help us on “how” to realize the operation. 3) select login 4) type in user name 5) type in password 6) submit

6. Conclusion and Future Work

A framework for goal definition has been proposed, after we constructed the goal-oriented system life cycle architecture.

By using this methodology, we could not only analyze the purpose of the proposed system, but also get the system definition to define a system for further implementation. We have provided a method for goal definition, which could be used in other goal-based methodologies.

In the future, specific methods will be proposed for generating system implementations, such as database, and system programs, to demonstrate the feasibility of our system definition.

References

- [1] Anton A. I., “Goal-Based Requirements Analysis”, Proceedings of the Second International Conference, Requirements Engineering, April 1996
- [2] Basili V., G. Caldiera and D. Rombach, “The Goal Question Metrics Approach”, Encyclopedia of Software Engineering, Wiley, 1994.
- [3] Betley L. D. and Whitten J. L., “System analysis and design for the global enterprise”, McGraw-Hill International Edition, Seventh Edition, 2007
- [4] Briand L. C., “An operational process for goal-driven definition of measures”, IEEE transactions on software engineering, VOL 28, No. 12, 2002
- [5] Chan S. L., “Information technology in business process”, Business Process Management Journal, Volume 6 Number 3, pp.224-237, 2000
- [6] Checkland, P., “Soft systems methodology: an overview”, J.Appl.Sys.Anal., 15, 27-30., 1988
- [7] Checkland, P., “Soft Systems Methodology: A 30-year Retrospective”, John Wiley & Sons, Ltd, Chichester., 1999
- [8] Dardenne A., A. van Lamsweerde, and S. Fickas, “Goal-directed requirements acquisition,” Science of Computer Programming, vol.20, pp.3–50, 1993.
- [9] Kavakli E., “Goal-oriented requirements engineering: A unifying framework,” Requirements Engineering, vol.6, pp.237–251, 2002.
- [10] Kelly P., Arguing Safety -A Systematic Approach to Managing Safety Cases, PhD Thesis, Department of Computer Science, University of York, 1998.
- [11] Kokune A., M. Mizuno, K. Kadoya, and S. Yamamoto, “FBCM: Strategy modeling method for the

- validation of software requirements,” *Journal of systems and software*, pp. 314-327, 2007.
- [12] Lee J, “Design rationale systems: understanding the issues”, *IEEE Expert: Intelligent Systems and their Applications*, pp.148, 1997
- [13] Mylopoulos John, Lawrence Chung and Brian Nixon, “Representing and using non-functional requirements: A process-oriented approach”, *IEEE Transactions on Software Engineering*, June 1992.
- [14] Olle T. W. et al., “Information system methodologies”, *International Federation for Information Processing*, Second Edition, 1991
- [15] Redwine S. and W. Riddle, “Software technology maturation,” *Proceeding 8th International Conference of Software Engineering*, IEEE CS Press, 1985, pp. 189–200.
- [16] Rolland C., G. Grosz, and R. Kla, “Experience with goal-scenario coupling in requirements engineering,” *Proc.RE’99*, pp.74–83, June 1999.
- [17] Web Dictionary, <http://en.wikipedia.org/>
- [18] Yu E.S.K., "Modeling organizations for information systems requirements engineering", *Proceedings of IEEE International Symposium on Requirements Engineering*, IEEE, 1993, 34-41.
- [19] Yu E. S. K., “Modeling organizations for information system requirements,” *IEEE International Symposium on Requirements Engineering*, pp.226-235, Jan. 1997

Automatic Semantic Annotation of Web Images

R.C.F. Wong

Abstract

As the number of web images is increasing at a rapid rate, searching them semantically presents a significant challenge. Many raw images are constantly uploaded with little meaningful annotation of semantic content, limiting their search and discovery. In this paper, we present a semantic annotation technique based on the use of image parametric dimensions and metadata. Using decision trees and rules induction, we develop a rule-based approach to formulate annotations and search for specific images fully automatically, so that by the use of our method, semantic query such as "sunset by the sea in autumn in New York" can be answered and indexed purely by machine. Experimental results indicate that this approach is able to deliver highly competent performance, attaining good recall and precision rates of sometimes over 80%. This approach enables a new degree of semantic richness to be automatically associated with images which previously can only be performed manually.

1 Introduction and Related Work

The number of web images is increasing at a rapid rate, and searching them semantically presents a significant challenge. Many raw images are constantly uploaded with little meaningful annotation of semantic content, severely limiting their search and discovery. While some sites encourage tags or keywords to be included manually, such is far from universal and applies to only a small proportion of images on the Web [13, 15, 17].

Research in image annotation has reflected the dichotomy inherent in the semantic gap, and is divided between two main categories: concept-based image retrieval and content-based image retrieval. The former focuses on retrieval by objects and high-level concepts, while the latter focuses on the low-level visual features of the image. Low-level visual features are indicated by visual content descriptors.

However, an advantage of using low-level features is that, unlike high level concepts, they do not incur any indexing cost as they can be extracted by automatic algorithms.

In contrast, direct extraction of high-level semantic content automatically is beyond the capability of current technology. Although there has been some effort in trying to relate low-level features such as blobs and regions [7] to higher-level perception, these are limited to isolated words, and they also require substantial training samples and statistical considerations [3–5, 8, 10].

Here, we present an automatic semantic image annotation technique based on a systematic analysis of image capture metadata in conjunction with image processing algorithms. The result is the ability to automatically formulate annotations to large numbers of web images which endows them with a new level of semantic richness. In doing so, we can answer semantic queries such as "Find images of sunset by the sea in New York in autumn" purely automatically for images without any form of manual involvement.

2 Correlating Scene Characteristics with Image Dimensions and Features

2.1 Scenes of Image

In relation to image acquisition, many images may be broken down to few basic scenes [14], such as nature and wildlife, portrait, landscape and sports. A landscape scene comprises the visible features of an area of land, including physical elements such as landforms, living elements of flora and fauna, abstract elements such as lighting and weather conditions. Landscape photography is the normal approach to ensure that as much of objects is in focus as possible, which commonly adopts a small aperture setting, since the smaller the aperture, the greater the depth of field in shots [2, 11]. In portrait photography, the goal of is to capture the likeness of a person or a small group of people. Like other types of portraiture, the focus of acquisition is the person's face, although the entire body and the background may be included [11]. Sports photography corresponds to the genre of photography that covers all types of sports. The equipment used by a professional photographer usually includes a fast telephoto lens and a camera that has an extremely fast exposure time that can rapidly take pictures. Definite relationships exist between the type



(a) night scenes

(b) outdoor portraits

(c) day scenes

(d) wildlife

(e) sports

Figure 1. Each column shows the top matches to semantic queries of (from left to right): "night scenes", "portraits", "day scenes", "wildlife" and "sports".

Table 1. scenes of images

Categories	Scenes	Symbols
Landscape	Day scenes	(S_d)
	Night scenes	(S_n)
	Sunrises and sunsets	(S_{ss})
Portraits	Indoor events	(S_{ie})
	Indoor portraits	(S_{ip})
	Outdoor events	(S_{oe})
	Outdoor portraits	(S_{op})
	Sports	(S_s)
Nature	Macro	(S_m)
	Wildlife	(S_w)

of scenes and image acquisition parameters. Some typical scene categories are given in Table 1, and scene images are given in Fig. 1.

An image I_i may be represented by a number of dimensions d_{i1}, \dots, d_{ik} which correspond to the image acquisition parameters.

$$I_i = (d_{i1}, \dots, d_{ik}) \quad (1)$$

Each dimension has a certain domain D_j , i.e. $d_{ij} \in D_j$, for all i . Each image corresponds to a point in k -dimensional space. Fig. 2 shows the image points of the images from Fig. 1. As we see from Fig. 2, each particular type of image scenes tend to cluster together, which forms the basis of our rule-induction algorithm in the next section. Each of these dimension values may be a scalar or vector. An example of scalar dimension is the exposure value (EV) [14] defined by:

$$EV = \log_2 \frac{f^2}{t} \quad (2)$$

where f is the relative aperture (f-number) and t is the exposure time.

An example of a vector-valued dimension is the GPS coordinate. For example, the GPS coordinates of Melrose Ave, Los Angeles is $\{34.083517, -118.321951\}$ in vector-valued format, which can be interpreted as $\{N34^\circ 5' 0.7'', W118^\circ 19' 19.0''\}$.

The image file format standard, embedded in the images and established by the Japan Electronic Industry Development Association [16], makes use of the Exchangeable Image File Format (EXIF) and contains metadata specification for image file format used in digital cameras. The specification uses the existing JPEG, TIFF Rev. 6.0, and RIFF WAVE file formats, with the addition of specific metadata tags. The Standard, defining image file system standards to enable image files to be exchanged among different recording media, was standardized in 1998 as a companion to the metadata standard. The most recent version, metadata standard version 2.2, was issued in 2004 with additional tag information and recording format options [16]. The metadata tags defined in the metadata standard cover a broad spectrum of data including: date and time information, acquisition parameters and descriptions and copyright information, which has shown to be useful for managing digital libraries [13], where limited tags and comments are entered manually. Some other common records of acquisition parameters including aperture (f), exposure time (t), subject distance (d) and focal length (L) and fire activation (h). Location information can be included in the metadata, which could come from a GPS receiver connected to the image acquisition devices. These GPS data are part of the metadata standard, are stored in a separate IFD within the metadata

of images, such data may be used to compute the physical location of images.

2.2 Rule-Induction and Annotation Algorithms

Here, we analyze the image dimensions to annotate and classify images. We annotate images with predefined semantic concepts in conjunction with methods and techniques of image processing and visual feature extraction. Our system structure is given in Fig. 3.

Our algorithm begins by constructing a decision tree starting from a training set, and each case specifies values for a collection of attributes and for a class. Our attributes includes discrete or continuous values. From [12], the information gain of an attribute a for a set of cases T is calculated as follows. If a is discrete, the T_1, \dots, T_s are the subsets of T consisting of cases with distinct value for attribute a , then:

$$gain = info(T) - \sum_{i=1}^s \frac{T_i}{T} \times info(T_i) \quad (3)$$

where

$$info(S) = - \sum_{j=1}^{N_{Class}} \frac{freq(C_j, S)}{|S|} \times \log_2 \left(\frac{freq(C_j, S)}{|S|} \right) \quad (4)$$

is the entropy function.

If a is a continuous attribute, cases in T are ordered with respect to the value of a . Assume that the ordered values are v_1, \dots, v_m . For $i \in [1, m - 1]$ the value

$$v = \frac{(v_i + v_i + v_{i+1})}{2}, \quad (5)$$

with splitting

$$T_1^v = \{v_j | v_j \leq v\}, T_2^v = \{v_j | v_j > v\}. \quad (6)$$

For each value v , the information gain $gain_v$ is computed by considering the splitting above.

The classifier c4.5 has the ability to induce annotation rules in the form of decision trees from a set of given examples. Firstly, we evaluate the best minimum number of branches for training subset m and testing subset n of the c4.5 classifier by comparing the respective average error rates E_m and E_n . Since at least two branches must contain a minimum number of objects and values over 50 could deliver least differentiated results, we alter the minimum number of branches, $2 < m \leq 50$ and $2 < n \leq 50$, in order to determinate the appropriate set of parameters.

We re-ran each parameter 3 times and averaged the results. Fig. 4a presents the results obtained with various m parameters. Although the value of E_m , where $m < 10$, for the training set is relatively low (around 0.05), E_n for

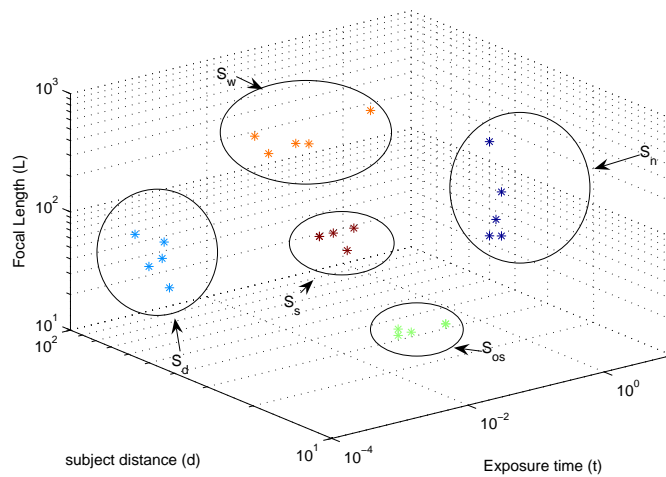


Figure 2. Image distribution in three-dimensional space

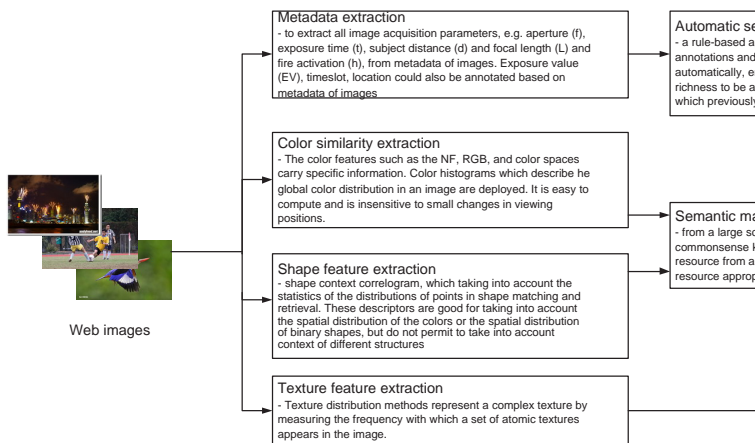


Figure 3. System structure

Table 3. Detailed experimental results of top annotation scenes of images

Scenes	Precision rate	Recall rate
(S_n)	91.6%	66.0%
(S_{op})	86.6%	59.6%
(S_d)	85.7%	89.8%
(S_m)	83.6%	84.5%
(S_w)	80.0%	29.2%
(S_s)	70.0%	38.3%
(S_{ip})	69.6%	63.2%

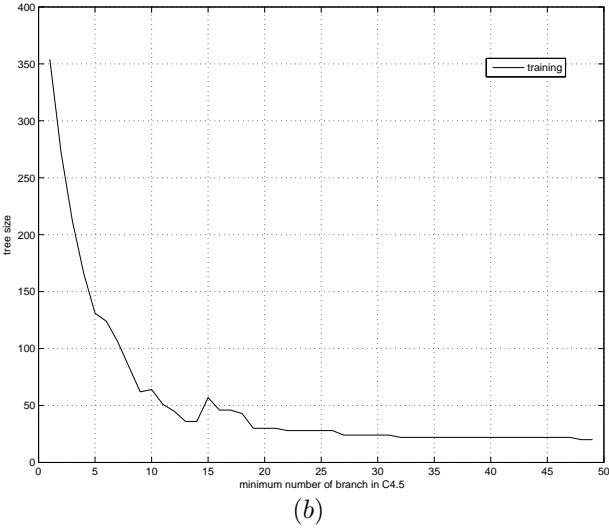
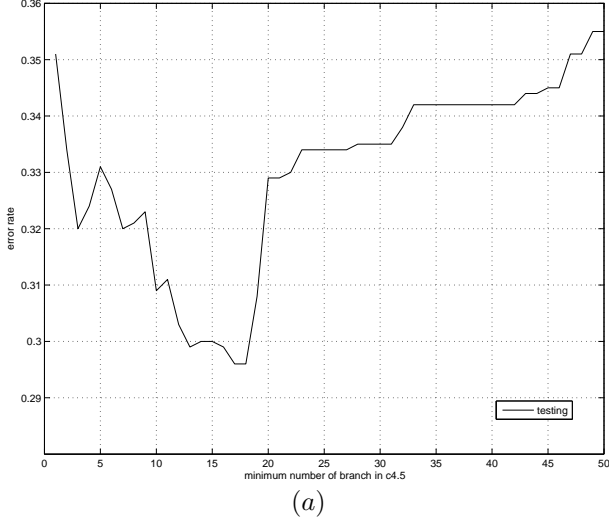


Figure 4. (a) Minimum number of branches of the c4.5 classifier (b) Tree size of the c4.5 classifier

$n < 10$ range from 0.3 to 0.35. When m and n fall between 15 and 20, E_n is least (0.3). Fig. 4b indicates that the size of the tree drops as the number of branches increases. The size of the tree drops sharply from 350 to 20 when m is increased from 2 to 25 and remains relatively constant when $m > 20$. In summary, $m = 17$ delivers the best score in terms of average errors and in tree size where $E_m = 0.274$, $E_n = 0.2944$. This produces the set of rules given below.

Let the set of images be I and $S_n, S_d, S_{ss}, S_{op}, S_{oe}, S_{ip}, S_{ie}, S_s, S_m, S_w \subset I$,

$$\forall i \in I, (t_i > 0.125) \wedge (d_i > 30) \wedge (EV_i \leq 8) \Rightarrow i \in S_n \quad (7)$$

$$\forall i \in I, (d_i > 30) \wedge (EV_i > 8) \wedge (t_i \leq 0.125) \Rightarrow i \in S_d \quad (8)$$

$$\forall i \in I, (f_i > 20) \wedge (d_i > 50) \wedge (EV_i > 11) \Rightarrow i \in S_{ss} \quad (9)$$

$$\forall i \in I, [(f_i \leq 5.6) \wedge (5 < d_i \leq 8)] \wedge \{(t_i \leq 0.00625) \wedge (L_i \leq 30)\} \vee [(30 < L_i \leq 182) \wedge (ISO_i \leq 250)] \vee (L_i > 182) \vee (t_i \leq 0.003125) \Rightarrow i \in S_{op} \quad (10)$$

$$\forall i \in I, (f_i > 5.6) \wedge (L_i \leq 25) \wedge (5 < d_i \leq 8) \wedge (t_i > 0.003125) \Rightarrow i \in S_{oe} \quad (11)$$

$$\forall i \in I, (f_i > 5.6) \wedge (0.003125 < t_i \leq 0.011111) \wedge (5 < d_i \leq 8) \wedge (L_i > 25) \Rightarrow i \in S_{ip} \quad (12)$$

$$\forall i \in I, (5 < d_i \leq 8) \wedge \{(f_i \leq 5.6) \wedge [(L_i \leq 30) \wedge (t_i > 0.00625)] \vee [(ISO_i > 250) \wedge (30 < L_i \leq 182)]\} \vee [(h_i = 1) \wedge (f_i > 5.6) \wedge (L_i > 25) \wedge (t_i < 0.011111)] \Rightarrow i \in S_{ie} \quad (13)$$

$$\forall i \in I, (d_i > 10) \wedge (150 < L_i \leq 400) \wedge (t_i \leq 0.005) \Rightarrow i \in S_s \quad (14)$$

$$\forall i \in I, (d_i \leq 5) \wedge (EV_i > 9) \Rightarrow i \in S_m \quad (15)$$

$$\forall i \in I, (L_i > 450) \wedge (d_i > 20) \Rightarrow i \in S_w \quad (16)$$

Our annotation system has been implemented using an image database of 3231 images downloaded from the Internet Web, which are stored in JPEG format with varying

Table 2. Number of images per scene category

dataset	S_d	S_n	S_{ss}	S_m	S_w	S_{ie}	S_{ip}	S_{oe}	S_{op}	S_s
training	92	26	7	71	19	201	209	120	490	53
testing	167	50	22	97	41	298	283	203	709	73
total	259	76	29	168	60	499	492	323	1199	126

sizes, ranging from 200×72 to $32,770 \times 43,521$ pixels. The file size of images are in between 22,397 to 768,918 bytes. All images were retrieved from one of the free photo album over the Web. Images are downloaded randomly and chosen without any bias. Except the metadata which embedded in the images, the system does not use any tag information in the matching process. We conduct experiments using those images to training and evaluate our approach using quantitative criteria. A summary of the properties of these datasets is given in Table 2.

Fig. 1 shows the top matches to semantic queries to the five types of scenes indicated at the bottom. Table 3 gives the detailed experimental results of top annotation scenes.

2.3 Features and Similarity Measures

In addition to the above annotation rules, we also enrich the annotation through color histogram matching. In order to explore the relationship between global color distribution and the scenes of images, MATLAB 7.0.4 is used. Firstly, we convert all images from RGB color space to an indexed image value and then extract the global color feature based on the color histogram. Histogram search is sensitive to intensity variation, color distortion and cropping. From [6, 9], Let I be an $n \times n$ image (for simplicity we assume that the image is square). The colors in I are quantized into m colors $c_1 \dots c_m$.

For a pixel $p = (x, y) \in I$, let $I(p)$ denote its color. Let $I_c \triangleq \{p | I(p) = c\}$. For convenience, we use the L_∞ -norm to measure the distance between pixels. i.e. for pixel $p_1 = (x_1, y_1), p_2 = (x_2, y_2)$, we define $|p_1 - p_2| \triangleq \max\{|x_1 - x_2|, |y_1 - y_2|\}$. we denote the set $\{1, 2, \dots, n\}$ by $[n]$

The *histogram* h of I for $i \in [m]$ is given by:

$$h_{c_i} \triangleq n^2 \cdot \Pr_{p \in I}[p \in I_{c_i}] \quad (17)$$

For any pixel in the image, $h_{c_i}(I)/n^2$ gives the probability that the color of the pixel is c_i .

The color distribution does not include any spatial information and this problem is especially acute for large database and robust to large appearance changes. We make use of the *correlogram* of I , which for $i, j \in [m], k \in [d]$ is given by:

$$\Upsilon_{C_i, C_j}^{(k)}(I) \triangleq \Pr_{p_1 \in I_{C_i}, p_2 \in I}[P_2 \in I_{C_j} | |p_1 - p_2| = k] \quad (18)$$

Given any pixel of color c_i in the image, $\Upsilon_{C_i, C_j}^{(k)}$ gives the probability that a pixel at distance k away from the given pixel is of color c_j . The *correlogram* of I captures spatial correlation between identical colors and measures the distribution of features such as color in the image as well as their spatial relationship [1, 9]. Another type of correlogram, the shape context correlogram [18], takes into account the statistics of the distributions of points in shape matching and retrieval. These descriptors are good for taking into account the spatial distribution of the colors or the spatial distribution of binary shapes.

After the image feature extraction process for all dataset, a scatter graph is plotted using the RGB indexed image value with their specified R, G and B values. This is shown in Fig. 6 and we see that images are almost centralized and cluster around axis $\{0,0,0\}$ and $\{1,1,1\}$. Only limited outliers are distributed on its surrounding. We believe that this is because object color of images tends to be evenly distributed unless specific scenes of images are intended to be presented, such as "night scenes by the sea" or "endless blue sky". Sunrises and sunsets are particularly suited to evaluate this as it offers a rich semantic meaning. Here, we select one image $i \in S_{ss}$ from training dataset with image indexed values $\{0.433333, 0.372549, 0.266666\}$. By covering all testing dataset of S_{ss} , indexed image values of all test dataset between $\{0.517647, 0.305882, 0.105882\}$ and $\{0.560784, 0.427451, 0.227451\}$ were evaluated. A total of 697 image were annotated which achieves a recall rate of 100% while the precision rate is only 1.3%. This indicates that by color alone, the annotation performance is not satisfactory. Fig. 7 shows that although the precision rate of the first dimension, global color distribution, is relatively low and delivers poor precision, using it in conjunction with the three other dimensions, the precision rate grows to 85.71%. Clearly, compared to annotation without the global color feature enabled, the performance is around 58.33% to 60%. In addition to the color, through the use of other measures, the shape and texture content may be similarly determined. In combination, therefore, a good level of semantic annotation accuracy can be achieved.

From the joint application of these techniques, we can








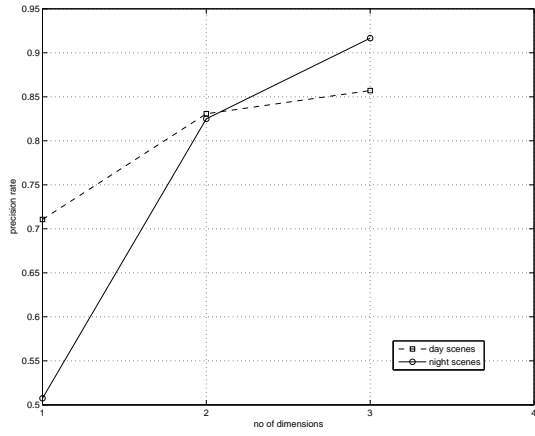
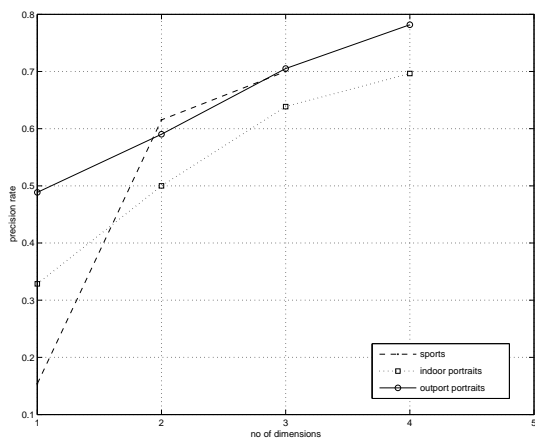
	<p>Landscape, night scenes, Victoria Harbor, Hong Kong, summer, night, sea, building</p>		<p>Portrait, indoor events, people, Cambridge, United Kingdom, spring, afternoon</p>
	<p>Nature, macro, animal, Taroko National Park, Japan, summer, morning leaf</p>		<p>Portrait, outdoor events, people, Cotton Tree Drive Marriage Registry, Hong Kong, autumn, afternoon</p>
	<p>Portrait, sports, people, Yio Chu Kang Stadium, Singapore, summer, afternoon, motion</p>		<p>Nature, wildlife, animal, Orlando Wetlands Park, Florida, United States, autumn, afternoon, feather, motion</p>
	<p>Landscape, day scenes, Chaopraya, Bangkok, Thailand, spring, morning, sea, building, sky</p>		<p>Portrait, indoor events, people, The Mesa Arts Center, Mesa, Arizona, United States, summer, night, motion</p>
	<p>Landscape, sunrise and sunset, SaiKung, Hong Kong, winter, evening, sea, sky, wood</p>		<p>Nature, wildlife, animal, Wetland Park, Hong Kong, autumn, afternoon, sea, feather</p>
	<p>Portrait, outdoor events, people, Yunlin County Stadium, Taiwan, winter, afternoon</p>		<p>Portrait, sports, people, Wulihe Stadium, Shenyang, China, summer, afternoon, motion</p>

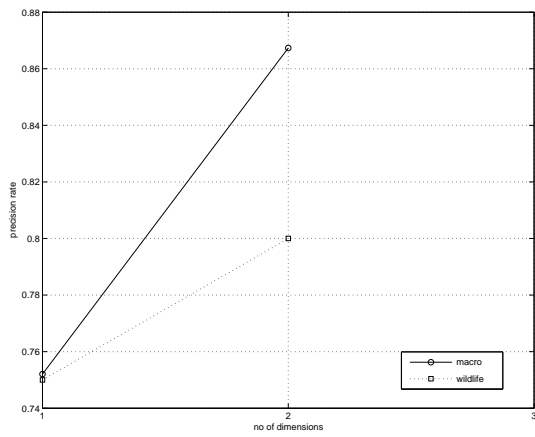
Figure 8. Semantic annotation of images



(a)



(b)



(c)

Figure 5. Precision rate results for automated annotation of scenes grouped by image categories (a) landscapes (b) portraits (c) natures

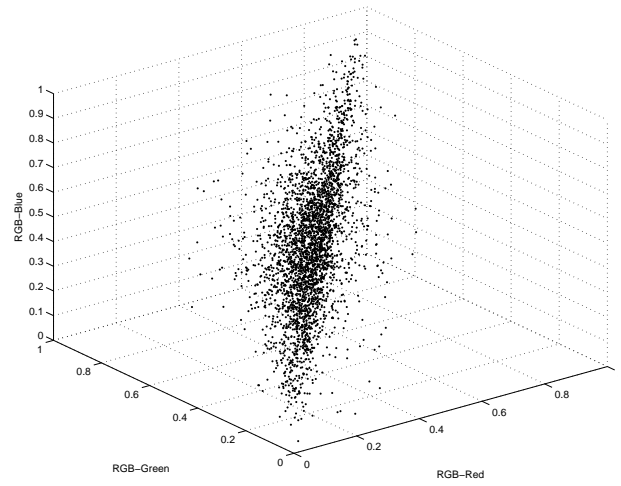


Figure 6. Clustering of color distribution

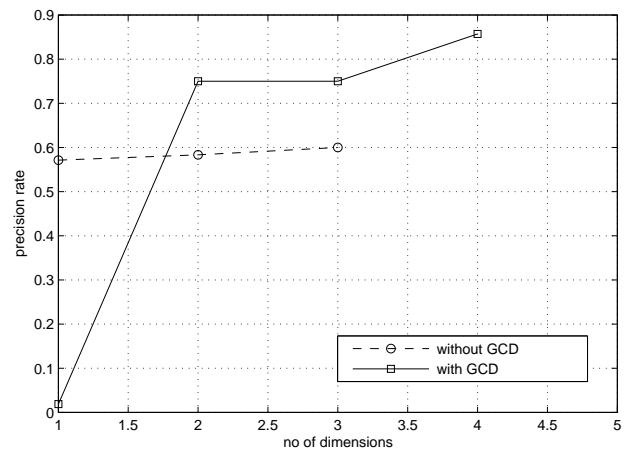


Figure 7. Incorporation of color feature extraction

formulate semantic annotations for specific image fully automatically and index images purely by machine without any human involvement. In Fig. 8, a sample set of images annotated semantically using the present approach is shown, which demonstrates the capability of the present method.

3 Conclusions

We have shown that by the systematic analysis of embedded image metadata and parametric dimensions, it is possible to determine the semantic content of images. Through the use of decision trees and rule induction, we have established a set of rules which allows the semantic contents of images to be identified. When jointly applied with feature extraction techniques, this produces a new level of meaningful image annotation. Using our image annotation method we are able to provide semantic annotation for any unlabeled images in a fully automated manner. Experimental results indicate that this approach is able to deliver highly competent performance, attaining good recall and precision rates of sometimes over 80%. This approach enables an advanced degree of semantic richness to be automatically associated with images which perviously can only be performed manually.

References

- [1] J. Amores, N. Sebe, and P. Radeva. Context-based object-class recognition and retrieval by generalised correlograms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1818–1833, October 2007.
- [2] T. Ang. *Dictionary of Photography and Digital Imaging: The Essential Reference for the Modern Photographer*. Amphoto Books, 2001.
- [3] K. Barnard, P. Duygulu, N. de Freitas, and D. Forsyth. *Exploiting text and image feature co-occurrence statistics in large datasets*. Chapter in Trends and Advances in Content-Based Image and Video Retrieval, Springer Lecture Notes in Computer Science Series, 2005.
- [4] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.
- [5] D. Blei, Michael, and M. I. Jordan. Modeling annotated data. *Proceedings of the ACM SIGIR Conference, Toronto*, pages 127–134, 2003.
- [6] S. Boughorbel, N. Boujemaa, and C. Vertan. Histogram-based color signatures for image indexing. *Information Processing and Management of Uncertainty in Knowledge-Based Systems, Annecy, France*, 3:977–984., 2002.
- [7] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. *Proceedings of the 7th European Conference on Computer Vision*, pages 97–112, 2002.
- [8] A. Frerri, G. Gallo, R. Giugno, and A. Pulvirenti. Best-match retrieval for structured images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:707–718, July 2001.
- [9] J. Huang, S. Kumar, M. Mitra, W. Zhu, and R. Zabih. Image indexing using color correlograms. *CVPR'97: IEEE Computer Vision and Pattern Recognition Conference*, pages 762–768, 1997.
- [10] M. Johnson, G. J. Brostow, J. Shotton, O. Arandjelovic, V. Kwatra, and R. Cipolla. Semantic photo synthesis. *Computer Graphics Forum*, 25(3):407–413, 2006.
- [11] R. Lenman. *The Oxford Companion to the Photograph*. Oxford University Press, 2005.
- [12] S. Ruggieri. Efficient C4.5. *IEEE Transactions Knowledge and Data Engineering*, 14(2):438–444, March 2000.
- [13] B. L. Saux and G. Amato. Image recognition for digital libraries. In *MIR '04: Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pages 91–98, New York, NY, USA, 2004. ACM Press.
- [14] R. F. Sidney. *Camera Exposure Determination*. In *The Manual of Photography*. Oxford: Focal Press, 9 edition, 2000.
- [15] Y. Sun, S. Shimada, and M. Morimoto. Visual pattern discovery using web images. *MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 127–136, 2006.
- [16] Technical Standardization Committee on AV and IT Storage Systems and Equipment and Standard of Japan Electronics and Information Technology Industries Association. *Exchangeable image file format for digital still cameras: Exif Version 2.2 JEITA CP-3451*, April 2002.
- [17] C. F. Tsai, K. McGarry, and J. Tait. Claire: A modular support vector image indexing and classification system. *ACM Trans. Information System*, 24(3):353–379, 2006.
- [18] T. Zöllner and J. M. Buhmann. Robust image segmentation using resampling and shape constraints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(7):1147–1164, July 2007.

Concept-based Multimedia Searching and Indexing

Alice W. S. Chan

Abstract

Technology on text mining and searching has developed well while the search on multimedia data extremely falls behind. Multimedia indexing is required to support the multimedia objects search based on its semantic concepts. In our indexing approach, each index has a score to reflect its importance to a multimedia object. User query result is based on the ranking of these scores, which can be changed over time due to the users' search behaviors. By analyzing the users' search behaviors, semantic concepts can be discovered and migrated through an index hierarchy. To provide better query result, the concept of genetic algorithm is also introduced.

Keywords: Dynamic Indexing, Genetic Algorithm, Semantic Concepts, User Feedback, Multimedia Search

1. Introduction

As the rapid growth of information technology and the Internet, multimedia information, such as visual images, audio and video files, has been a general type of information around us. In the WWW, any desired information is reachable by using search engines. Therefore, the need of multimedia search becomes significantly common. However, most of the search engines are only well developed in text-based web page searches.

In today's Internet, most of the search engines applied the PageRank algorithm for ranking query results based on scoring and the link structure of the Web [6], [7]. By applying this algorithm, the relative importance of hyperlinked set of documents can be measured. However, multimedia information search is far more difficult than searching text-based documents since the content of text-based documents can be extracted automatically and relatively easily while the content of multimedia objects can not and complicated.

Many different methods and techniques have been proposed for retrieving images [1], [2], [4], [5]. They are mainly classified into two main categories; "concept-based" image retrieval, and "content-based" image retrieval. The former focuses on using words to retrieve images (e.g. title, keywords, and caption), while the latter

focuses on the visual features of the image (e.g. size, colours, and textures). In an effective "concept-based" multimedia retrieval system, efficient and meaningful indexing is necessary [3]. Due to the current technology limitations, it is impossible to extract semantic content of the multimedia objects automatically. Meanwhile, the discovery and inclusion of new indexing terms is always costly and time-consuming. Therefore, some manual indexing is required for an initial multimedia searching system.

As the well development of web 2.0, it relies heavily on user-generated content and users' expert knowledge [8]. The Wikipedia is one of the successful examples of the web 2.0. In our proposed method, we adopted the spirit of the web 2.0 and proposed the collaborative indexing by the Internet users. Through the continuous and extensive use of the multimedia search system, users tend to provide more meaningful indexing and expert judgment for the system. Consequently, multimedia objects would be optimally indexed such that semantic visual information search on those objects would become possible.

2. Index Elements and Structure

In our study, we only focused on the indexing of semantic contents of multimedia objects and exclude the metadata. Since the indexing of metadata is relatively straightforward and less meaningful than the semantic contents that perceived by human. For example, indexing and retrieving a song by its characteristics (e.g. style) is more meaningful than indexing its metadata (e.g. track number).

2.1. Index Element

In the proposed multimedia system, we consider a set of data objects $\{O_j\}$. In addition, the characteristics or semantic contents of each element object O_j (multimedia data objects such as images, video, or music) in this set can not be extracted automatically. For every O_j , it links with a set of index I_j that consists of a number of elements:

$$I_j = \{e_{j1}, e_{j2}, \dots, e_{jMj}\}. \quad (1)$$

Each index element e is a triple, such that

$$e_{jk} = (t_{jk}, s_{jk}, o_j), \quad (2)$$

where t_{jk} is an index term ID, s_{jk} is the score associated with t_{jk} , and o_j is the object ID. The higher the score s_{jk} , the index term t_{jk} is more important to the object o_j . The relationship of t_{jk} , s_{jk} , and o_j can be represented in the following entity relationship diagram (ERD).

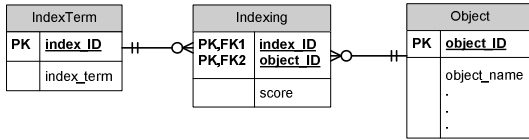


Fig.1. ERD on t_{jk} , s_{jk} , and o_j

2.2. Index Hierarchy

The index hierarchy refers to the index sets of all the objects stored in the database. By partitioning the value of score s_{jk} , it can be divided into N levels L_1, L_2, \dots, L_N with using a set of parameters P_1, P_2, \dots, P_N . (See Fig. 2.)

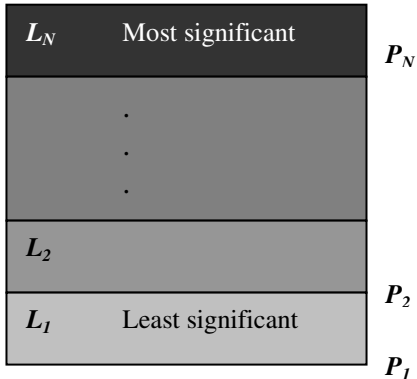


Fig. 2. Partitioning of the Index Set

By considering score value x of a given index term, the index term would be placed in level L_i if

$$P_i \leq x < P_{i+1}, \text{ where } i = 1, \dots, N-1, \quad (3)$$

and would be placed in level N if

$$P_i \leq x. \quad (4)$$

In this index hierarchy, the higher the level, the more significant than those in the lower levels to an object. Hence, multimedia data would be search in the top level first. In some case, we may neglect the lower levels.

2.3. Minimal Index

When an object O is *minimally indexed*, it means:

- (i) O has only a single index term T , and
- (ii) where T is a single word.

Since the multimedia objects are searched by search term(s) in a search query, the objects should be indexed in order to be searched. Therefore, it is necessary to add a minimal index to a multimedia object when user adding objects to the database. The *unindexed* multimedia object, which is not indexed by any index term, could not be found.

2.4. Growth of Index

Consider an object K that is minimal indexed with a term T_1 . K can be searched by user query which contains T_1 . It may have chance that many objects returned in the query result since many objects are indexed with T_1 . Among these returned objects, user can distinguish objects by adding another index term T_2 to K . Thus, user can search the desired object by entering both index terms in the search query.

For example, when we consider the searching of a song “Für Elise”, which is a piece of music composed by Beethoven. Initially, we assume that the audio object is minimally indexed with the term “Beethoven”. User can search this song by the term “Beethoven”. Sometimes, some user query would be more specific, with both search terms “Beethoven” and “Für Elise” are used. Same multimedia object would be returned in the result when searching by the term “Beethoven”, since the term “Für Elise” is not indexed yet. Eventually, user would select the audio object “Für Elise” and suggest a new index term, “Für Elise”, to this music. Thus, the new index term would be included in the low level of the index hierarchy for this audio object. For every query that having both terms, “Beethoven” and “Für Elise”, user would select this audio object and increase the score on the index terms for that object. Thus, the score of the index would be increased and the new index would be proper installed. (The increase of score will be introduced in the next section.) Through progressive usage, the indexing on multimedia objects would be enriched.

3. Score Updating Algorithms

User search behaviors, such as result selection and the relevance feedback, would affect the score directly. By continuously use of the search system, more user search behaviors can be collected and analyzed. The following

will introduce how the scores are affected by the user search behaviors.

In this section, we will consider an example on a user input search query $Q(T_1, T_2)$, N multimedia objects O_1, O_2, \dots, O_n are returned in the query result and ordered by the corresponding score S_1, S_2, \dots, S_n in descending order.

3.1. Score Increment

By considering this example, the related index scores on T_1 and T_2 for the desired object O_x would be increased by the following cases:

- (i) When user select O_x in the query result list, or
- (ii) When user provide positive feedback on O_x

In these two cases, the related index scores on T_1 and T_2 for the desired object O_x would be increased by a predefined value Δ_x , where the predefined value for these two cases can be different.

3.1. Score Decrement

By considering this example, there are two cases that would cause the index score decrease:

- (i) When user provide negative feedback on O_x , the related index scores on T_1 and T_2 for the desired object O_x would be decreased, or
- (ii) When user do not click on any object on the query result list, the related index scores on T_1 and T_2 for all objects O_1, O_2, \dots, O_n in the query result would be decreased.

In these two cases, the score would be decreased by a predefined value Δ_y , where the predefined value for these two cases can be different.

4. Object Ranking

When user input a series of search terms T_1, T_2, \dots, T_n in a search query $Q(T_1, T_2, \dots, T_n)$, the query score $S(Q|O_j)$ for an object o_j can be extracted by the following SQL statement:

```
SELECT Score
FROM Indexing
WHERE Object_ID = "Oj"
and Index_ID = "Ti-ID";
```

This score implies the relative importance of an index term T_i to the corresponding object O_j . Thus, the

multimedia objects in the query result should be ordered by score in descending order. The query result can be obtained by the following SQL statement:

```
SELECT Object_ID, SUM(Score) AS s
FROM Indexing
WHERE Index_ID in (T1-ID, T2-ID, ..., Tn-ID)
GROUP BY Object_ID
ORDER BY s DESC
```

Usually, users are only interested in the search results that appearing on the first few pages since the top ranked objects are more relevant to the search query than the low ranked objects in general. For this reason, users may miss some good choice of search result in some situations and the higher scored objects always have higher chance of score increase.

When we consider a huge number (e.g. 1000+) of multimedia objects returned by a query result, user may not reach their desired object since the object always ranked very low and nearly "hidden". We proposed to apply the genetic algorithm (GA) in discovering those "hidden" objects. By the randomness chrematistics of GA, it helps those objects to promote to a higher ranking position eventually.

With the GA, the object ranking in the query result is determined by a probability value:

$$P_i = \frac{\sum_{j=1}^i S_j}{\sum_{j=1}^n S_j}, \text{ where } i = 1, 2, \dots, n. \quad (5)$$

The higher the probability value would have greater chance of getting higher rank in the query result list.

5. Similar Work

YouTube is a video sharing website where users can upload, view and share video clips [9]. Users are allowed to add tags (similar as the index term) to the videos, such that user can search videos by words.

6. Conclusion and Future Works

In this paper, we introduced the general idea of concept-based multimedia searching and dynamic indexing. We proposed an index hierarchy that relating multimedia objects with index terms while the relevant importance is reflected by the corresponding score, which evolves continuously through users' extensive use of the system.

Currently, we found that users tend to be passive in giving relevant feedback. Although users has higher

chance of giving feedback in the age of web 2.0, it is impossible for users to provide feedback for every multimedia objects for a search result. Furthermore, the number of index terms of a multimedia should be enriched to improve the search performance. However, it is costly for adding index terms manually.

In the future works, we will consider the improvement on the relevance feedback mechanism and the way of automatic index enrichment. Also, we will discover the relationship of user search behaviors and the index hierarchy.

7. References

- [1] Azzam, I., Leung, C. H. C., and J. Horwood: "A fuzzy expert system for concept-based image indexing and retrieval", *Proceedings of the IEEE International Conference on Multi-media Modeling*, Melbourne, Australia, 2005, pp. 452-457.
- [2] Azzam, I., Leung, C. H. C., and J. Horwood: "Implicit concept-based image indexing and retrieval", *Proceedings of the IEEE International Conference on Multi-media Modeling*, Brisbane, Australia, January 2004, pp. 354-359.
- [3] Jaime Gomez and Jose Luis Vicedo: "Next-Generation Multimedia Database Retrieval", *IEEE Multimedia*, July 2007, pp. 106-107
- [4] Leung C. and Liu J.: "Multimedia data mining and searching through dynamic index evolution", *Advances in Visual Information systems, 9th International Conference, VISUAL 2007*, Shanghai, China, June 2007, pp.298-309.
- [5] Over, P., Leung, C. H. C., Ip, H., and M. Grubinger: "Multimedia retrieval benchmarks", *IEEE Multimedia*, Vol. 11, No.2, 2004, pp. 80-84.
- [6] Taher H. Haveliwala: "Topic-sensitive PageRank: a context-sensitive ranking algorithm for web search", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 15, No. 4, 2003, pp. 784-796.
- [7] Wikipedia, "PageRank", Retrieved December 26, 2007 from the WWW: <http://en.wikipedia.org/wiki/Pagerank>
- [8] Wikipedia, "Web 2.0", Retrieved December 26, 2007 from the WWW: http://en.wikipedia.org/wiki/Web_2.0
- [9] Wikipedia, "YouTube", Retrieved December 26, 2007 from the WWW: <http://en.wikipedia.org/wiki/Youtube>

Path Planning of Virtual Human by Reinforcement Learning

Yuesheng He

Abstract

Virtual Human can hold the possibility of performing a variety of assistive and analysis tasks in 3D virtual environments. However, widespread use of avatars assistants in these environments requires ease of use by individuals who are generally not skilled on designing operators. In this paper we present a method of training Virtual Humans that bridges the gap between user designing and building of a virtual human's action and autonomous learning of a avatar task. With our approach to variable autonomy, we integrate reinforcement ability at the level of virtual human's special regime action into for learner to permit faster policy acquisition. We illustrate the ideas using a virtual human's environment of tasks that planning a path across a plane and a pyramid with a numbers of obstacles to reach a certain target and get good result

1. Introduction

In Virtual human, the representation of the geometric and behavioral characters of human in the virtual environment^{[1][2][3]}, is one of the new research areas of computer science.

In the industry area, virtual human can effectively support the ergonomics (human factor) and training of operation. It can be used in each periods of the digital product management including design, producing, maintaining and training^{[2][3]}

One of the interesting area on Virtual Human is to make it more automatically or even more intelligent. The more automatic the Virtual Human is, the more time people can save. In the same time the animation which has been created shall be more like the reality^[1]

For the applications of virtual human, such as simulating tasks of human's action in the buildings or cities, or accomplish a certain job in a certain environment, the requirement is further amplified by the fact that the user is generally not a skilled engineer and can therefore not be expected to be able or willing to provide constant, detailed instructions^{[19][22]}.

In fact, the synthesis of human locomotion has always been a challenging problem in computer animation. Numerous studies from varies fields, such as biomechanics, robotics, and ergonomics, have provided a

rich data base on "normal" straight-walk gait patterns. However, the capability of walking over uneven terrain and cluttered environments is fundamental in our daily life and critical on some occasions, such as exploring new environments^[14]

So, the capability of walking in the 3D environment with obstacles is an important function of computer animation system for virtual human.

Ideally, motion control mechanisms of human walking simulation should be^[7]:

- Broadly capable: they should not be limited to periodic gaits along simple paths on even terrain, but adaptable to the environment. Also a variety of walking modes and styles should be possible;
- Easily controlled: the user should have convenient, hierarchical control over the motions. At the high level, ideally, through a small number of intuitive parameters, the system should be able to generate the corresponding walking motion. At the low level, additional locomotion attributes are provided to simulate a variety of walks;
- Responsive: they should generate motions in minimal latency response to user inputs. This capability is important in helping the animator to direct the desired motions, and also critical in virtual environment applications;
- Realistic: they should be able to generate natural human walking motions.

Thus, to accomplish the walking tack automatically is a key function of virtual human which should be carefully considered and designed.

In this paper, a method of path planning is described to support real-time creation of human walking in virtual environments.

The motion control technique integrates studies from animation, biomechanics, human gait experiments, and psychology, and represents an important initial step toward meeting the locomotion requirements in diverse environments.

First, any high level plan for the virtual human must be on the base of the low level motion control and can support any optimization approaches which have possibility to be integrated into the motion control

mechanism to simulate walking in different environments. Second, the method should give virtual human “online” planning ability to adapt different virtual environment. Finally, it is responsive. Since relatively simple inverse kinematics mechanisms and optimal search algorithms are widely used in the computation, interactivity can be easily achieved, which would make the method well suited for virtual environment applications.

To achieve the requirement, a reinforcement learning method is presented in the paper. The second section will give the description of it. The Third section will give some experimental result. Then, the fourth section will give a discussion and our future work.

2. Methodology of Virtual Human’s path planning

2.1 The framework

Each presentation In this section we introduce our reinforcement learning methodology for the virtual human’s path planning.

This means that an avatar in the virtual environment can automatically learn a behavioral model. We have developed a new technique to perform the learning behavior to achieve the foundational requirements which have been presented in section one.

As the movement and postures should be used by the planning level, the low level is to solve the problem of how to describe the elemental action of a virtual human. So, it is described a finite automata as below^[10]

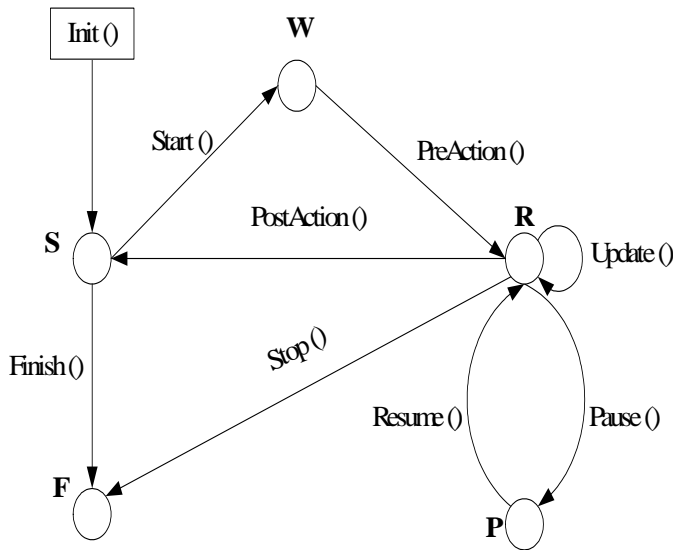


Figure 1 States of action finite automata

The 5 different States are:

1. S -- Stopped : accomplished the initialization
2. W--Waiting: ready to start the simulating loop
3. R --Running: running the simulating loop
4. P --Paused: simulating loop is paused
5. F --Finished: the action is finished

A planning based reinforcement learning technique will only learn a sub-optimal policy, the quality of which depends on the search depth limit. To ensure that on the whole an avatar's behavior is optimal, we utilize the Learning approach and dynamic programming.

To use the property of Q-learning to accomplish the task, we have developed an alternative approach to support the learning model by computing discrete examples of a policy.

To treat the action as a discrete model, the planning level can be described as a tree search as below:

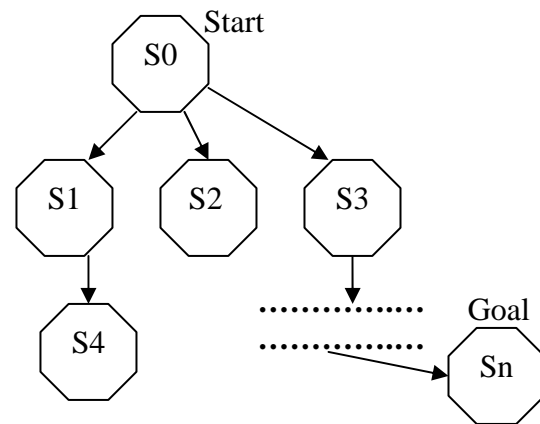


Figure 2 Searching space of the reaction with the environment by treating the action is discrete

Every state in the figure 2 is based on the finite automata which has been presented by figure 1.

The sum of the avatar Q-values for a particular state determines the optimal action^{[8][7]}. This requires each avatar to indicate, from its perspective, a value for every action. For instance, the avatar i reports its action values $Q_i(s, a)$ for the current state s to the arbitrator. Then, the arbitrator chooses an action maximizing the sum of the Q-values. In our proposed methodology we use distance value as a pseudo values to represent the Q-values.

In fact, the low and high level model of action of Virtual Human can not only describe the former examples, but the other ones.

The approximate idea of the reward function which, when optimized, should generate a “desirable” behavior. However the key part of reinforcement learning of planning Virtual Human’s action is to represent a barrier to the broader applicability of the optimal control function and design the algorithms.

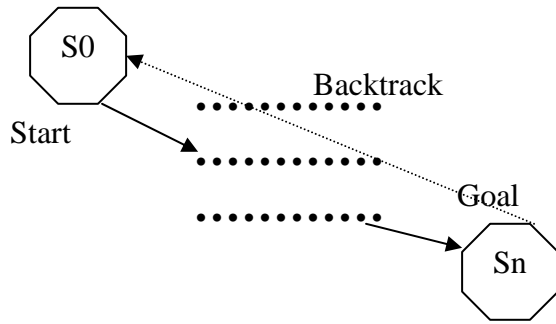


Figure 3 Basic model of reward of actions

As the figure 3 shows, the model can “backtrack” the reward of a certain state then the value of the whole chain of the states will be renewed. Thus, the sum of the Q-value should be updated. So the structure has the ability to support the reinforcement learning.

2.2 MDP

From the very beginning of this section, we recall the definition of a Markov Decision Process (MDP) (Puterman, 1994). A (finite) MDP is defined by the tuple (X, A, R, P) , where X and A denotes the finite set of states and actions, respectively. $P : X \times A \times X \rightarrow [0, 1]$ is called the transition function, since $P(x, a, y)$ gives the probability of arriving at state y after executing action a in state x . Finally, $R : X \times A \times X \rightarrow \mathcal{R}$ is the reward function, $R(x, a, y)$ gives the immediate reward for the transition (x, a, y) ^[12]

The process of the path planning of Virtual Human in a virtual environment or dynamic situations can be treated as a Markov Decision Process.

Since the ultimate goal of decision making is to find an optimal behavior subject to some optimality criterion. Optimizing for the infinite-horizon expected discounted total reward is one of the most important parts of such criteria. Under this criterion, we are trying to find a policy that maximizes the expected

value of $\sum_{t=0}^{\infty} \gamma^t \cdot r_t$, the r_t is the immediate reward in the step t and the $0 \leq \gamma < 1$ is the discount factor.

A basic way to find out an optimal policy to calculate the optimal value function: $V : X \rightarrow \mathcal{R}$, that gives the values of each state.

Formally, the optimal value function should be satisfy the below recursive equation known as Bellman equations (Bellman, 1957):

$$V(x) = \max_a \sum_y P(x, a, y)(R(x, a, y) + \gamma V(y))$$

all $x \in X$

Besides the state value function, some other types of value function can be defined as well. One typical example is the state –action-value function $Q : X \times A \rightarrow \mathcal{R}$, which satisfies the formula:

$$Q(x, a) = \sum_y p(x, a, y)(R(x, a, y) + \gamma \max_{a'} Q(y, a'))$$

all $x \in X$

In this case, $Q(x,a)$ has the meaning – the expected value of taking action a in state x while following the optimal policy. In this way, the value of state action pairs are learned instead of state values, which enable model-free learning (Watkins, 1989). The learning algorithm that under this theory is called Q-learning . Moreover, the value which is optimizing by the learning process is the “Q-value”^[12]

2.3 Q-Learning

The characteristic of reinforcement learning is a trial-and error feature. The reward will be given when the answer to a question is correct, while the penalty will be awarded when there is an error^[13]. There are three elements involve in reinforcement learning of virtual human in the virtual environment, they are:

- environment state;
- records of the information which impacts an action;
- action which is performed to direct to the state.

The relationship between Virtual Human and virtual environment is as the figure below:

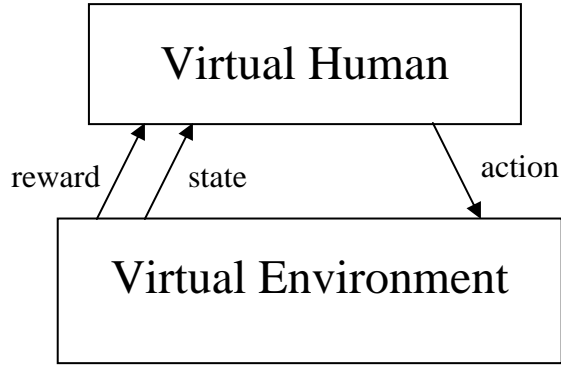


Figure 4 The relationship between Virtual Human and Environment

To allow learning work be successful, one can set the goal with the elements of Reward function, which indicates the value of action in each state and the Value function indicating the total reward in each state.

In the work presented here, user commands at a high level of instructions are presented to the virtual human in 3 dimensional graphical virtual environment goal point to be reached. Then the virtual human perform certain actions (walk or run) or suggested specific actions to execute. This input is used, as long as it conforms with the a priori constraints, to temporarily drive the avatar. At the same time, user commands play the role of training input to the learning component, which optimizes the autonomous control policy for the current task. Here, Q-learning (Watkins,1989) is used to estimate the utility function, $Q(s,a)$, by updating its value when action a is executed from state s according to the formula^[13]:

$$Q(s,a) \leftarrow Q(s,a) + \alpha(r + \gamma \max_{a'} Q(s',a') - Q(s,a))$$

In the formula, the “r” is the reward and γ is the discount factor.

The total reward used by the Q-learning algorithm throughout virtual human operation is:

$$r = r_g + r_w$$

In this case the r_g represents the avatar can reach the goal or not. The r_w represents the weight of the sub factors of actions such as the distance or terrain.

So the whole process is treated as a MDP and used the reinforcement learning (in this case, it is Q-Learning) to recursively find the optimal solution.

3. Experimental Result

We chose to simulate the method of path planning based on the Q-Learning on the Matlab. Since the Virtual human is move in a 3 dimensional space, the simulating environment also adopt the assumption that the working space is 3 D.

The first diagram represents the 3 D obstacles. The second diagram represents the initial place with a square, the goal with a x and obstacles with the contours. Then the third diagram represents the planning path with a red line.

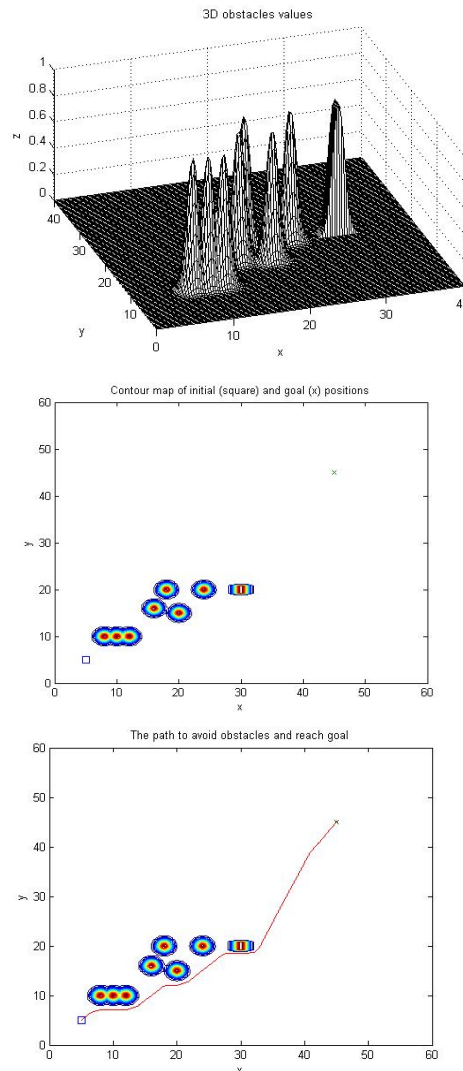


Figure 5 Simulation of the Path Planning in the 3 dimensional environment

If the positions of the obstacles have been changed, the method can adapt the change and recompute the optimal path under the criteria of the same reward of Q function.

The figure below shows a simulating result under the different environment which has different obstacles:

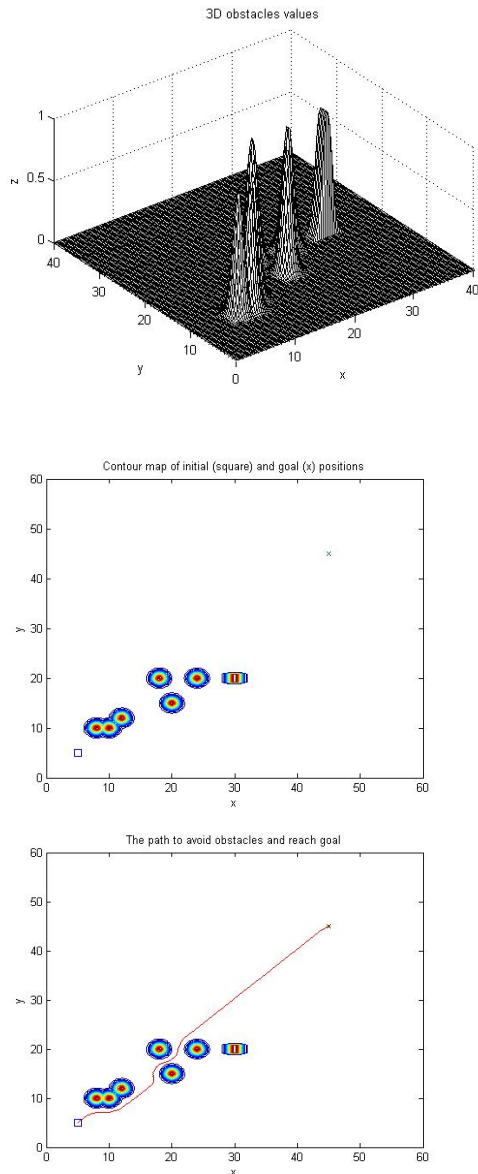


Figure 6 Simulation of path planning by changing the position of obstacles

We can test our behavioral control method of Virtual Human by implementing the method in C++ and Python thanks to the open source code “ReplicantBody” on the virtual human simulations.

ReplicantBody is a character animation toolkit written in C++, built upon Cal3D, ConfigScript and OpenSceneGraph Features.

Copyright (C) 2003 VRlab, Ume University.

It is distributed under GNU LGPL.

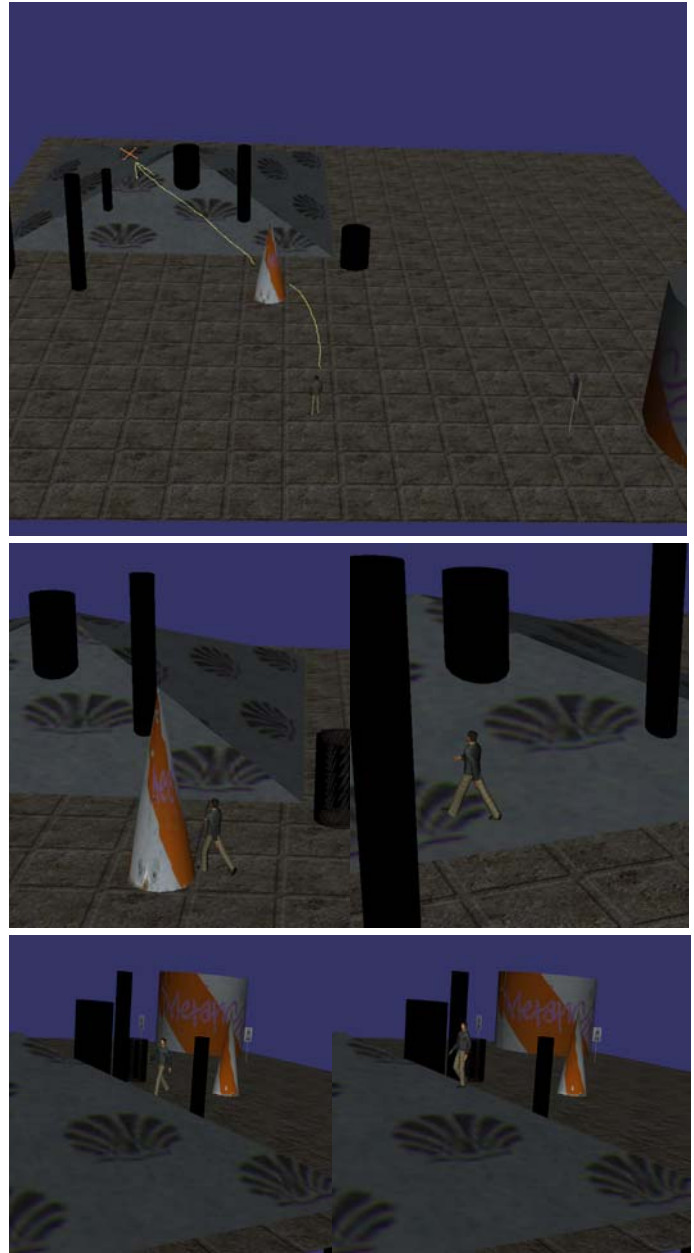


Figure 7 Example of Virtual Human walks in the Virtual Environment with Obstacles

As we can see, the method can control the virtual human to walk naturally and plan an optimal path to cross 3 dimensional obstacles to reach the goal. Moreover, when the system of animation is running, the frame rate is more than 50 frames per second. That means it has ability to support real time animation.

4. Conclusion and Discussion

Our reinforcement learning method of path planning for virtual human has proved to be fast, simple and adaptable. It has also succeeded in automatically learning behavioral movement for different situation. Thus, we believe it will be useful to the virtual human animation system of simulating human actions.

Because of the method has been integrated with the low level behavioral model of the avatar well, it can well support the natural behavior of the virtual human such as walk and run.

Because the reinforcement learning does not need the supervisor signal, but only the reward from the environment, the virtual human gets the ability to adapt different circumstance.

Also because the framework of the method has the ability to fit the levels structure of the description of Virtual Human's behavioral model. The virtual human under this method is responsive, realistic and easy to control.

However Q-learning is sometimes impractical because of a large Q-factor table^{[6][7]}. If the environment is much more complicated than the experimental environment, the result will be not so ideal. On the other hand, though the method has ability to compute dynamic states, the computational complication will be higher. Thus multi virtual humans in a same environment, the requirement of computational resources will be very high.

Our future work is to simplify the model of the Q-learning to make it to support much more complicated environment. Otherwise, we are going to analyze the multi virtual human's behavior more deeply to find out more efficient learning algorithm to control their actions.

References

- [1] Norman I Badler. Virtual humans for animation, ergonomics, and simulation[J]. Nonrigid and Articulated Motion Workshop, 1997. Proceedings. IEEE Published: 1997, Page(s): 28 -36.
- [2] Norman I. Badler, C. B. Phillips, B. L. Webber. Simulating Humans: Computer Graphics, Animation, and Control. Oxford University Press, 1993.
- [3] Norman I. Badler, Jan Allbeck, Liwei zhao, Meeran Byun. Representing and Parameterizing Agent Behaviors. In Proceedings of Computer Animation'02, 133-143, 2002.
- [4] Li Yan, Wang Wei, Lu Xiaojun, An Anthropometry-based Method for Virtual Human Modeling. Journal of System Simulation, 15(supplementary issue): 210-212, 2003.
- [5] T. Conde, D. Thalmann, Learnable Behavioural Model for Autonomous Virtual Agents : Low-Level Learning, In Proceedings of Fith International Conference on Autonomous Agents and Multiagent Systems 2006 (AAMAS-06), Hakodate, Japan, May 2006, pp. 89-96
- [6] T. Conde, D. Thalmann, An Integrated Perception for Autonomous Virtual Agents: Active and Predictive Perception, Computer Animation and Virtual Worlds, Volume 17, Issue 3-4, John Wiley, 2006
- [7] Moccozet L, Thalmann N. M., Dirichlet Free-Form Deformation and their Application to Hand Simulation[J], Proceedings Computer Animation'97, IEEE Computer Society, 1997, pp.93-102.
- [8] Molet T, Boulic R, Thalmann D. A real-time anatomical converter for human motion capture[J]. In Euro graphics Workshop on Computer Animation and Simulation, 1996, 79-94.
- [9] He Yuesheng, Li Yan, Lu Xiaojun, A Software Architecture of a Virtual Human Factors Analysis System for Maintenance Engineering, Computer Simulation, Issue4,2006
- [10] Xiaojun Lu, Yan Li, Hangen He, Yunxiang Ling: Research and Implement of Virtual Human's Walking Model in Maintenance Simulation. Edutainment 2006: 1027-1036
- [11] N. Magnenat-Thalmann, A. Foni, G. Papagiannakis, N. Cadi-Yazli. Real Time Animation and Illumination in Ancient Roman Sites. The International Journal of Virtual Reality, IPI Press, Vol.6, No.1, pp. 11-24. March 2007.
- [12] István Szita, B'álint Tak'acs deim, Andr'as L'orincz, ϵ - MDPs: Learning in Varying Environments. Journal of Machine Learning Research 3 (2002) 145-174
- [13] Theodore J. Perkins PERKINS, Andrew G. Barto BARTO, Lyapunov Design for Safe Reinforcement Learning, Journal of Machine Learning Research 3 (2002) 803-832
- [14] Vladim'ir Step'an, Jir' Z'ara, V'aclav Hlav'ac, Presenting generalized human activities in virtual environment, 2005 ACM 1-59593-203-6/05/0005
- [15] Alejandra Garc'ia Rojas M., Fr'ed'eric Vexo, Daniel Thalmann, Individualized Reaction Movements For Virtual Humans, GRAPHITE 2006,2006 ACM 1-59593- 64- /06/0011, 79-85

- [16] Robert C. Hubal, Geoffrey A. Frank, Curry I. Guinn, Lessons Learned in Modeling Schizophrenic and Depressed Responsive Virtual Humans for Training, IUI'03,2003 ACM 1-58113-586-6/03/0001, 85-92
- [17] Ronan Boulic, Pascal BCcheiraz, Luc Emering, Daniel Thalmann, Integration of Motion Control Techniques for Virtual Human and Avatar Real-Time Animation, ACM VRST '97, 111-117
- [18] Weixiong Zhang, Randall W. Hill, Jr., A Template-Based and Pattern-Driven Approach to Situation Awareness and Assessment in Virtual Humans, Agents 2000, ACM 2000 1-58113-230-1/00/6, 116-123
- [19] Z.M. Ruttkay, D. Reidsma, A. Nijholt, Human Computing, Virtual Humans and Artificial Imperfection, ICMI'06, ACM 1-59593-541-X/06/0011, 179-184.
- [20] Karl Tuyls, Katja Verbeeck, Tom Lenaerts, A Selection-Mutation Model for Q-learning in Multi-Agent Systems, AAMAS'03, ACM 1-58113-683-8/03/0007, 693-700
- [21] Catherine Zambakal, Amy Ulinski, Paula Goolkasian, Larry F. Hodges, Social Responses to Virtual Humans: Implications for Future Interface Design, CHI 2007 Proceedings, ACM 978-1-59593-593-9/07/0004, 1561-1570
- [22] Edward M. Sims, Reusable, lifelike virtual humans for mentoring and role-playing, Computers & Education 49 (2007) 75–92
- [23] Lucio Ieronutti , Luca Chittaro, Employing virtual humans for education and training in X3D/VRML worlds, Computers & Education 49 (2007) 93–109

Tensor Locality Preserving Projections for Face Recognition

Limin Cui

Abstract

Automated face detection and recognition is one of the most attentional branches of biometrics and it is also the one of the most active and challenging tasks for computer vision and pattern recognition. Over the past few years, some embedding methods have been proposed for feature extraction and dimensionality reduction in various machine learning and pattern classification tasks. Locality Preserving Projection (LPP) has been used in such applications as face recognition and image. In this paper, we propose some novel tensor embedding methods which, unlike previous methods, take data directly in the form of tensors of arbitrary order as input. These methods allow the relationships between dimensions of a tensor representation to be efficiently characterized. Extensive experiments show that our methods are not only more effective but also more efficient.

1. Introduction

Real data are often very high-dimensional in pattern recognition and computer vision. The problem of dimensionality reduction appears in many fields of data mining, machine learning, and computer vision. It is a necessary preprocessing step in the face recognition system for simplification of the data and noise reduction. The goal of dimensionality reduction is to map the high-dimensional samples to a lower dimensional space such that certain properties are preserved. Furthermore, the discovered low-dimensional structures can be used for classification, clustering, and data visualization. Traditional dimension reduction techniques such as Principal Component Analysis (PCA) [1], factor analysis can not discover nonlinear structures embedded in the set of data points.

As a new nonlinear dimensionality reduction approach, manifold learning has attracted the attention of many researchers. A manifold is a topological space with locally Euclidean. Manifolds offer a powerful framework for dimension reduction. The key idea of dimension reduction is to find the most succinct low dimensional structure that is embedded in a higher dimensional space. The major algorithms include Isometric mapping (ISOMAP) [2], Locally Linear Embedding (LLE) [3], Laplacian Eigenmaps [4], Hessian Eigenmaps [5],

Diffusion maps [6], and so on. The approach can be used for discovering the intrinsic dimensions of nonlinear high-dimensional data effectively and aim researchers to analyze the data better.

Most of previous works on face recognition represent a face image by a vector in high-dimensional space. For example, if a face image of size is $n_1 \times n_2$ pixels, we represent it by a vector in $\mathbb{R}^{n_1 \times n_2}$ and face space denote the set of all the face images. The face space is usually a low dimensional manifold embedded in the ambient space. The curse of dimensionality is suffered for ignoring the underlying data structure.

On the other hand, real data in our processing are often multidimensional and naturally represented as 2nd-order (matrices) or higher order tensors. For example, a face image is intrinsically 2nd-order tensor, and a video sequence is 3rd-order tensor. Many researchers focus on multilinear algebra to represent data in their natural form. The typical linear algorithms for learning such a face manifold for recognition include PCA, Linear Discriminant Analysis (LDA) [7] and Locality Preserving Projection (LPP) [8].

Multiple factors make face recognition to become more difficult such as illumination, viewpoint, viewing direction, pose, camera characteristics, and so on. M. Alex O. Vasilescu and Demetri Terzopoulos [9-11] take advantage of multilinear algebra, the algebra of higher-order tensors, to obtain a parsimonious representation that separates the various constituent factors. Their new representation of facial images, called TensorFaces. They use Tensorface to represent the set of face images by a higher order tensor and extend Singular Value Decomposition (SVD) to higher order tensor data. In this way, the multiple factors related to expression, illumination and pose can be separated from different dimensions of the tensor, such as Fig. 1.

Locality Preserving Projections (LPP) are linear projective maps that arise by solving a variational problem that optimally preserves the neighborhood structure of the data set. When the high dimensional data lies on a low dimensional manifold embedded in the ambient space, LPP are obtained by finding the optimal linear approximations to the eigenfunctions of the Laplace Beltrami operator on the manifold. Although LPP are linear projections, they have many good properties to be similar with nonlinear techniques such as Laplacian

Eigenmaps or Locally Linear Embedding. But LPP only process vectorized data, it also suffer from the curse of dimensionality and the high computation. In this paper, a new multilinear approach to face recognition — Tensor Locality Preserving Projections (TLPP) algorithm is proposed. TLLP is a natural extension of LPP to the multilinear case. It is particularly useful in applications when the data samples are naturally represented as matrices or higher-order tensors.

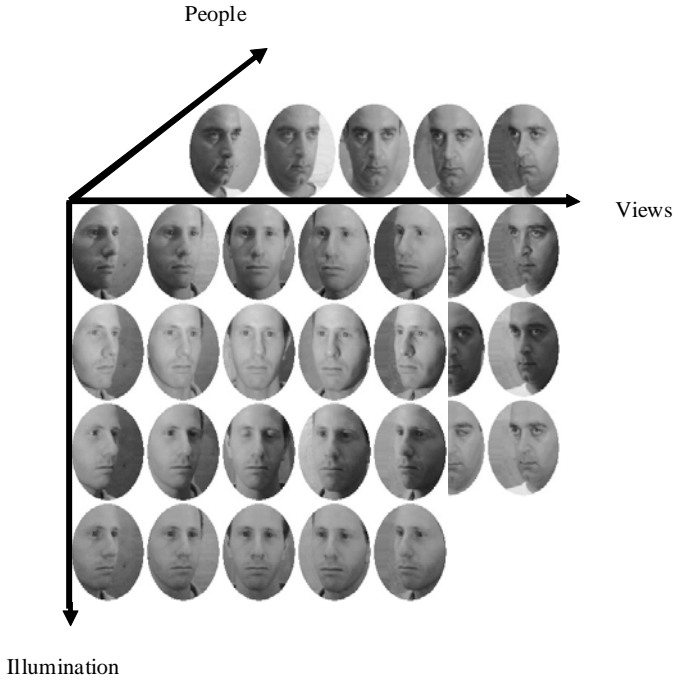


Fig.1 The multiple factors of Weizmann face image database

The paper is organized as follows. Section 2 and 3 review LPP and Multilinear approaches. The proposed method is introduced in section 4. Experimental results are presented in section 5. Finally, conclusions are given in section 6.

2. Locality Preserving Projections (LPP)

LPP is a linear dimensionality reduction algorithm, and firstly builds a graph incorporating neighborhood information of the data set. As we known, LPP is an optimal linear approximation to Laplacian Eigenmap. Using the notion of the Laplacian of the graph, we can compute a transformation matrix which maps the data points to a subspace. Although LPP are linear transformation, they can optimally preserve local neighborhood information in a certain sense. The algorithm is described as follows

Given n data points $x_1, \dots, x_n \in \mathbb{R}^m$, LPP seek a transformation matrix K that maps each data point x_i to a corresponding lower-dimensional data point y_i , where $y_i = K^T x_i$. Different criteria are used in different methods for finding the transformation matrix.

Step 1: Construct the adjacency graph G using ε neighborhoods or k nearest neighbors.

Step 2: Choose the weights w_{ij} of two edges according to the adjacency graph G . There are two methods to compute w_{ij} , namely

$$w_{ij} = e^{-\frac{\|x_i - x_j\|^2}{t}} \quad \text{or} \quad w_{ij} = 1,$$

when nodes i and j are connected. Then we can obtain W is a sparse matrix.

Step 3: Compute the eigenvectors and eigenvalues for the generalized eigenvector problem:

$$XLX^T a = \lambda DX^T a$$

where D is a diagonal matrix whose entries are column or row sums of W , $D_{ii} = \sum_j w_{ji}$, $L = D - W$ is the

laplacian matrix. The optimization problem for LPP is given by

$$\arg \min_a a^T XLX^T a \quad \text{s.t.} \quad a^T XLX^T a = 1$$

3. Multilinear Approach

The notation and basic definitions of multilinear algebra will be introduced in this section.

Tensors are multilinear mappings over a set of vector spaces. For example, a tensor is a higher order generalization of a vector (first order tensor) and a matrix (second order tensor).

An k -th-order tensor is denoted as: $A \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_k}$. The r -mode product of a tensor A and a matrix $U \in \mathbb{R}^{J_r \times I_r}$ is an $I_1 \times I_2 \times \dots \times I_{r-1} \times J_r \times I_{r+1} \times \dots \times I_k$ tensor denoted by $A \times_r U$ whose entries are

$$(A \times_r U)_{i_1 \times i_2 \times \dots \times i_{r-1} \times j_r \times i_{r+1} \times \dots \times i_k} = \sum_{i_r} A_{i_1 \times i_2 \times \dots \times i_{r-1} \times i_r \times i_{r+1} \times \dots \times i_k} U_{j_r i_r}$$

The scalar product $\langle A, B \rangle$ of two tensors $A, B \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_k}$ is defined as

$$\langle A, B \rangle = \sum_{i_1} \dots \sum_{i_k} A_{i_1 \dots i_k} B_{i_1 \dots i_k}$$

The Frobenius norm of a tensor A is defined as $\|A\| = \sqrt{\langle A, A \rangle}$.

In general, the goal of linear dimensionality reduction

in a tensor space can be described as follows. Given n data points $A_1, \dots, A_n \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_k}$, a linear tensor embedding method seeks to find k transformation matrices $U_i \in \mathbb{R}^{I_i \times I_i}$, where $I_i < I_i, i=1,2,\dots,k$, such that n corresponding embedded data points $B_1, \dots, B_n \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_k}$ can be obtained as

$$B_i = A_i \times_1 U_1 \times_2 U_2 \times \dots \times_k U_k$$

Where $i=1,2,\dots,n$.

According to tensor definition, the mode- r vectors of a k th order tensor A are the I_r -dimensional vectors obtained from A by varying index i_r while keeping the other indices fixed. The mode- r vectors are the column vectors of matrix $A_{(r)} \in \mathbb{R}^{I_r \times (I_1 \dots I_{r-1} I_{r+1} \dots I_k)}$ that results by mode- r flattening the tensor A . Figure 2 is described a flattening 3rd-order tensor. The tensor can be flattened in 3 ways to obtain matrices comprising its mode-1, mode-2, and mode-3 vectors.

4. Tensor Locality Preserving Projections

Given data points A_1, \dots, A_n from an unknown manifold M embedded in a tensor space $\mathbb{R}^{I_1 \times I_2 \times \dots \times I_k}$, firstly we construct a neighborhood graph G to represent the local geometric of M . The corresponding affinity matrix $S = [s_{ij}]_{n \times n}$ is defined based on the heat kernel as:

$$s_{ij} = \begin{cases} e^{-\frac{\|A_i - A_j\|^2}{t}}, & \text{if } A_i \in O(K, A_j), \text{ or } A_j \in O(K, A_i) \\ 0, & \text{otherwise} \end{cases}$$

where $O(K, A_i)$ denotes the set of K nearest neighbors.

Let $U_i \in \mathbb{R}^{I_i \times I_i}$, where $i=1,2,\dots,N$, be the corresponding transformation matrices. Based on the neighborhood graph G , the optimization problem for TLPP can be expressed as:

$$\arg \min Q(U_1, \dots, U_N) = \sum_{i,j} \|A_i \times_1 \dots \times_N U_N - A_j \times_1 \dots \times_N U_N\|^2 s_{ij}$$

$$\text{s.t. } \sum_i \|A_i \times_1 \dots \times_N U_N\|^2 d_{ii} = 1$$

where $d_{ii} = \sum_j s_{ij}$ is, the more important is the data points B_i in the embedded tensor space for representing the data point A_i . It is easy to see that the objective function will give a high penalty if neighboring points A_i and A_j are mapped far apart. Thus if two points A_i and

A_j are close to each other, then the corresponding points B_i and B_j in the embedded tensor space are also expected to be close to each other. We solve this optimization problem by applying an iterative scheme. Assuming that $U_1 \times U_2 \times \dots \times U_{r-1} \times U_{r+1} \times \dots \times U_k$ are known, we denote $y_i^r = A_i \times_1 U_1 \times \dots \times_{r-1} U_{r-1} \times_{r+1} U_{r+1} \times \dots \times_k U_k$. Based on the properties of tensor and trace, we reformulate the optimization function in (7) as follows:

$$\arg \min P_r(U_r) = \text{tr} \left\{ U_r \left(\sum_{i,j} (Y_i^{(r)} - Y_j^{(r)}) (Y_i^{(r)} - Y_j^{(r)})^T s_{ij} \right) U_r^T \right\}$$

$$\text{s.t. } \text{tr} \left\{ U_r \left(\sum_i Y_i^{(r)} (Y_i^{(r)})^T d_{ii} \right) U_r^T \right\} = 1$$

The unknown transformation matrix U_r can be obtained by solving for the eigenvectors corresponding to the r_n smallest eigenvalues in the generalized eigenvalue equation as the follows

$$\left(\sum_{i,j} (Y_i^{(r)} - Y_j^{(r)}) (Y_i^{(r)} - Y_j^{(r)})^T s_{ij} \right) u = \lambda \left(\sum_i Y_i^{(r)} (Y_i^{(r)})^T d_{ii} \right) u$$

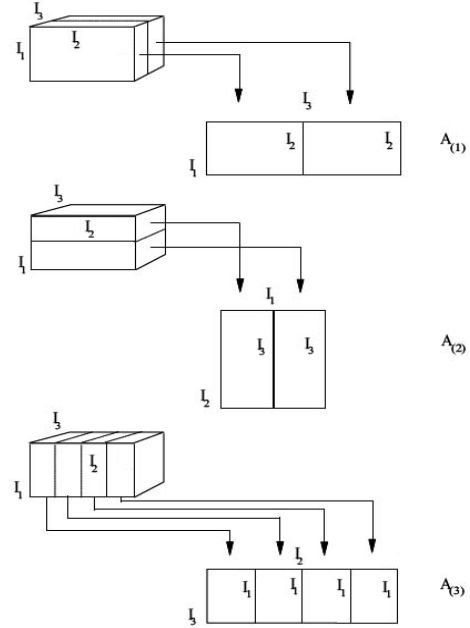


Fig.2 Flattening a 3rd-order tensor.

5. Experiments

As we stated earlier, image formation depends on scene geometry, viewpoint, and illumination conditions.

Multilinear algebra offers a natural approach to the analysis of the multifactor structure of image ensembles and to addressing the difficult problem of disentangling the constituent factors or modes.

In this section, several experiments are carried out to show the efficiency and effectiveness of our proposed algorithm for face recognition. We compare our algorithm with the PCA, LDA, and LPP, three of the most popular linear methods for face recognition.

The experiment is performed on the Cambridge ORL face database, which contains 40 distinct persons as shown in Fig. 2 Each person has ten different images. There are many variations in facial expressions such as open or closed eyes, smiling or nonsmiling, and glasses or no glasses. All the images were taken against a dark homogeneous background with the subjects in an up-right, frontal position, with tolerance for some side movements. There are also some variations in scale. We show two individuals as shown in Fig. 3.

In our face recognition experiments on the ORL database, we select 200 samples (5 for each individual) randomly as the training set. The remaining 200 samples are used as the test set. Using our method, we can obtain high recognition rate as shown in TABLE I.



Figure 3 Forty distinct persons in ORL face database.



Figure 4 Four individuals in the ORL face database. There are 10 images for each person.

Table 1. Experiment result on ORL face database

Methods	PCA	LDA	LPP	TLPP
Accurate Rate	85.7	96.3	96.6	97.1

6 Conclusions

Based on some recently proposed embedding methods, we have developed generalizations which can take data directly in the form of tensors of arbitrary order as input. Not only do our methods inherit the attractive characteristics of the previous methods in terms of exploiting the intrinsic local geometric and topological properties of the manifold, they are also appealing in terms of significant reduction in both space complexity and time complexity. Face recognition experiments based on the ORL database demonstrate that our tensor embedding methods give very impressive results.

References

- [1] Turk M. A. and Pentland A. P., “Eigenfaces for recognition”, *Journal of Cognitive Neuroscience*, Vol. 3, No. 1, pp. 71-86, Mar. 1991.
- [2] J.B. Tenenbaum, V. de Silva, and J. C. Langford , “A Global Geometric Framework for Nonlinear Dimensionality Reduction” *Science* 22 December 2000: Vol. 290. no. 5500, pp. 2319 – 2323.
- [3] T. S. Roweis and L. K. Saul, “Nonlinear Dimensionality Reduction by Locally Linear Embedding “ *Science* 22 December 2000: Vol. 290. no. 5500, pp. 2323 – 2326.
- [4] M. Belkin and P. Niyogi. “Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering”. *Advances in Neural Information Processing Systems* 15. Vancouver, British Columbia, Canada, 2001.
- [5] D.L. Donoho and C. Grimes “Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data”, *Proceedings of the National Academy of Science*, vol. 100, Issue 10, p.5591-5596, 2003.
- [6] Ronald R Coifman and Stephane Lafon and Ann Lee and Mauro Maggioni and Boaz Nadler and Frederick Warner and Steven Zucker, “Geometric

Diffusions as a tool for Harmonic Analysis and structure definition of data. Part I: Diffusion maps”, Proc. of Nat. Acad. Sci. no. 102 (2005), pp. 7426—7431.

- [7] Belhumeur P. N., Hespanha J. P., and Kriegman D. J., “Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, No. 7, pp. 711-720, July 1997.
- [8] Xiaofei He, and Partha Niyogi, “Locality Preserving Projections”, Advances in Neural Information Processing Systems 16, Vancouver, Canada, 2003.
- [9] M. A. O. Vasilescu, D. Terzopoulos, "Multilinear Subspace Analysis for Image Ensembles," Proc. Computer Vision and Pattern Recognition Conf. (CVPR '03), Vol.2, Madison, WI, June, 2003, 93-99.
- [10] M. A. O. Vasilescu, D. Terzopoulos, "Multilinear Image Analysis for Facial Recognition," Proceedings of International Conference on Pattern Recognition (ICPR 2002), Vol. 2, Quebec City, Canada, Aug, 2002, 511-514.
- [11] M. A. O. Vasilescu, D. Terzopoulos, "Multilinear Analysis of Image Ensembles: TensorFaces," Proc. 7th European Conference on Computer Vision (ECCV'02), Copenhagen, Denmark, May, 2002, in Computer Vision -- ECCV 2002, Lecture Notes in Computer Science, Vol. 2350, A. Heyden et al. (Eds.), Springer-Verlag, Berlin, 2002, 447-460.

Moving Object Detection Based on Information Theoretic Spatio-Temporal Saliency

Chang Liu

Abstract

This paper proposes to employ the visual saliency for moving object detection via direct analysis from video. Object saliency is represented by an Information Saliency Map (ISM), which is calculated from spatio-temporal patches. We use dimensionality reduction and kernel density estimation to develop an efficient information theoretic based procedure for constructing the ISM. It is shown that the ISM can be used to successfully detect foreground objects with different moving speeds and under different illumination variations. Two publicly available visual surveillance databases namely CAVIAR and PETS are selected for evaluation. Experimental results show that the proposed method is robust for both fast and slow moving object detection under illumination changes. The average detection rate is 93.92% while the false detection rate is 2.16% in CAVIAR (INRIA entrance hall and shopping center) dataset with ground truth data. Moreover, the detection rate is about 8 fps with resolution 320×240 in a typical Pentium IV computer.

1. Introduction

Moving object detection from video is the first and important step in video analysis [4] [24] [21] [9]. It is widely used as low-level tasks of computer vision applications such as target tracking, visual surveillance, human behavior recognition, video retrieval and a pre-stage of MPEG4 image compression. The objective of moving object detection is to locate different types of moving objects in the scene such as humans, cars, bicycles for further processing. The challenge of this research is to detect objects with different moving speeds in complex background clusters, and under different illumination changes. To tackle the problems, a number of motion detection methods have been proposed in the last decade. We loosely categorize these methods into three approaches, namely region-based approach, orientation-based approach and contour-based approach.

In region-based approach, background subtraction is the most popular method to detect moving objects. The rationale of this approach is to build an appropriate representa-

tion (background image model) of the scene so that the object(s) in the current frame can be detected by subtracting the current frame with the background image [20]. Based on this idea, several adaptive background models have been proposed. Stauffer and Grimson [18] developed a method to model each pixel as a Mixture Of Gaussians (MOG) and constructed a model that can be updated on-line. Along this line, other similar methods have been developed [25] [23]. But a common problem in background subtraction is that it requires a long time for estimating the background image model. Furthermore, because MOG assumes all pixels are independent, pixel correlation is not considered, so the background model based on individual pixels is sensitive to illumination and noise. When the density function is more complex and cannot be estimated parametrically, a non-parametric approach is more suitable to model background. Elgammal et al [5] proposed a set of Gaussian kernels for modeling the density at pixel level. This model estimates the probability density function (PDF) directly from the data without assumptions of the underlying distributions. Mittal et al [14] proposed an adaptive kernel density estimation (KDE) based background subtraction method and introduced a new bandwidth function which is data-dependent for density estimation. Normalized features are used to solve illumination problems. This method claimed to be able to handle mild illumination effects. More recent work on nonparametric background modeling can be found in [12] [15]. Generally speaking, moving object detection methods based on background model requires an accurate estimation of the background image. The performance is good if the background image does not change much in a short period of time.

In orientation-based approach, optical flow is the most widely used method. This approach approximates the motion of objects by estimating vectors originating or terminating at pixels in image sequences, so it represents the velocity field which warps one image into another high dimensional feature space. Based on optical flow technique [19] [17], these methods can accurately detect motion in the direction of intensity gradient, but the motion which is tangential to the intensity gradient cannot be well represented by the feature map. Moreover, optical flow based methods

also suffer from the illumination problem.

In the contour-based approach, level sets [2], active contours [22] and geodesic active contours [6] have been proposed. These methods can effectively detect moving objects with different sizes and shapes, and claimed to be insensitive to illumination changes [7]. But contour-based methods cannot handle the fast moving objects very well and is computationally expensive.

This paper proposes a new saliency-based approach for moving object detection from video. The idea of saliency [11] [16] has been employed for landmark and object detection from image. Unlike existing method, a new spatio-temporal model which incorporate both spacial and temporal saliency is proposed for moving object detection via direct analysis from video. Object saliency is represented by an Information Saliency Map (ISM), which is calculated from spatio-temporal patches. We use dimensionality reduction and kernel density estimation to develop an efficient information theoretic based procedure for constructing the ISM. Experimental results show that the ISM can be used to successfully detect foreground objects with different moving speeds under different illumination variations.

The rest of this paper is organized as follows. Section 2 will report the details of spatio-temporal ISM construction and its advantages in detecting moving object with different moving speeds and under different illumination variations. Experimental results and the conclusion are given in Sections 3 and 4 respectively.

2. Proposed Method

This paper proposes a new method to construct an information saliency map using both spacial and temporal information saliency. The temporal information saliency provides effective information to detect moving object, but suffer from two limitations. First it is sensitive to the illumination changes. Second, results on slow moving object may not be good. To solve the first limitation, Lambertian model is employed. To solve the second limitation, spacial information saliency is used as it is independent on the motion. The proposes method combines the temporal information saliency and spacial information saliency, and generates an information saliency map for moving object detection. The ISM is a two dimensional matrix and each entry reflects the spatio-temporal saliency of the corresponding pixels in that video frame. Therefore, by analyzing the ISM, the detection of moving object(s) with different motion speed under different illuminations is feasible.

2.1. Information theory

From modern attention theory, saliency is the impetus for selective attention. Different attention models may give different definitions of saliency [10]. In this paper, we use the *information* measure as a quantity that reflects saliency

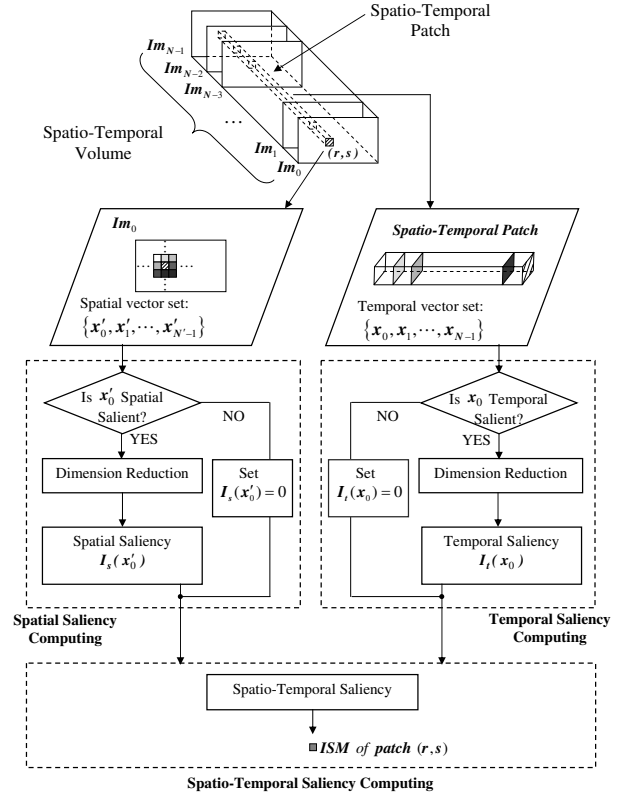


Figure 1. Flowchart of Information Saliency Map (ISM) computing from one single patch. It mainly contains three parts, namely spatial saliency computing, temporal saliency computing and spatio-temporal saliency computing

[3]. This measure is calculated as self-information in a particular context. Consider a discrete random variable $X \in 1, \dots, K$, suppose event $X = k$ is observed, the Shannon's self-information content of this event $I(k)$ is defined as follows,

$$I(k) = \log_2 1/p(X = k) = -\log_2 p(X = k) \quad (1)$$

It means that the information content of an event k is inversely proportional to the probability of the observation of event k . A rarely happen event contains high information while an event which happens frequently contains low information. The property of information theory shows close relationship with saliency, and information theory can be seen as a channel between saliency and selective attention [3].

2.2. Information Saliency Map (ISM)

We construct the information saliency map (ISM) based on spatial saliency and temporal saliency. For each frame in the video, we compute its ISM which shows the visual saliency of each pixel of the frame.

Figure 1 shows the block diagram of the proposed method in calculating the ISM. Our method consists of three steps, namely spatial saliency computing, temporal saliency computing and spatiotemporal saliency computing. Suppose we want to calculate the ISM for the frame Im_0 , a spatiotemporal (ST) 3D volume is constructed by the current frame Im_0 and its previous (N-1) frames, i.e. $Im_1, Im_2, \dots, Im_{N-1}$. The ST volume is then divided into smaller ST sub-volumes with smaller size of $M \times M \times N$. For each sub-volume, a spatial vector set $X' = (x'_0, x'_1, \dots, x'_{N'-1})$ is constructed by the patch (x'_0) in frame Im_0 and its $N' - 1$ neighborhoods. The spatial saliency is then computed. For the temporal saliency, a temporal vector set with N elements $X = (x_0, x_1, \dots, x_{N-1})$ is constructed from the sub-volume. It is noted that x'_0 is the same as x_0 . The temporal saliency will be calculated based on the temporal vector set. By combining the spatial and temporal saliency, the spatiotemporal saliency for x_0 is then determined. Similarly, the ISM for all other patches are computed. The ISM for the current frame Im_0 can be obtained as follows,

$$I(Im_0) = \begin{pmatrix} I(1,1) & I(1,2) & \dots & I(1,w) \\ I(2,1) & I(2,2) & \dots & I(2,w) \\ \vdots & \vdots & \ddots & \vdots \\ I(h,1) & I(h,2) & \dots & I(h,w) \end{pmatrix} \quad (2)$$

where $I(r, s)$ is the spatio-temporal information saliency for patch (r,s) , $r = \{1, 2, \dots, h\}$, $s = \{1, 2, \dots, w\}$.

2.2.1 Computing Temporal Saliency

This section computes the temporal saliency. Theoretically, based on the entropy equation, the temporal saliency of the sub-volume $I_t(r, s)$, $r = \{1, 2, \dots, h\}$, $s = \{1, 2, \dots, w\}$, is computed by the following equation:

$$\begin{aligned} I_t(x_0) &= I_t(r, s) \\ &= -\log_2[P(Im_0(r, s)|V(r, s))] \\ &= -\log_2(P(x_0|X)) \end{aligned} \quad (3)$$

where $V(r, s)$ represents the sub-volume constructed at $Im_0(r, s)$, x_0 is the vector form of $Im_0(r, s)$. After calculating all the sub-volumes in the frame Im_0 , the temporal saliency map $I_t(Im_0)$ for frame Im_0 is given by integration of $I_t(r, s)$, with the same structure to Im_0 in Eq.(2).

Calculating Eq.(3) for every patch is time consuming. Very often, many sub-volumes may not be temporally salient. To reduce the complexity of the proposed method, an effective verification method using vector variance ($\sum_{i=1}^N (x_i - \bar{x})^2$) is employed to determine whether a sub-volume is a temporally salient volume (TSV) or not. When there is no object motion or background change, the variance of the sub-volume will be very small. In that case, the sub-volume is not temporal saliency and the temporal

saliency value $I_t(r, s)$ can be set to zero. If the variance is larger than a threshold, we have to estimate the conditional probability in Eq.(3).

To compute the probability in Eq.(3), we need to find its distribution density function. Estimating the distribution in high dimension space $X = \{x_0, x_1, \dots, x_{N-1}\}$ is time consuming and requires many samples. To solve this difficulty, principal component analysis is employed for dimension reduction and only the first q principal component vectors $Y = \{y_1, y_2, \dots, y_q\}$ are used for estimating the probability. Therefore we have

$$I_t(x_0) = -\log_2(P(y_0|Y)) \quad (4)$$

we adopt a non-parametric approach and kernel density estimation (KDE) is employed. KDE obtains the exact probabilities regardless of the shape of the population distribution from which the random samples are drawn.

Considering the q -dimensional sample space $Y = \{y_1, y_2, \dots, y_N\}$, the multivariate kernel estimator is adopted and defined as:

$$\hat{f}(y) = \frac{1}{N} \sum_{i=1}^N K_H(y - y_i) \quad (5)$$

where the kernel $K_H(y) = \|H\|^{-1/2} K(H^{-1/2}y)$, H is the bandwidth matrix which specifies the spread of the kernel around sample y_i . In this paper, we use the sample-point estimator [14]:

$$\begin{aligned} \hat{f}(y) &= \frac{1}{N} \sum_{i=1}^N K_{H(y_i)}(y - y_i) \\ &= \frac{1}{N} \sum_{i=1}^N \|H(y_i)\|^{-1/2} K(H(y_i)^{-1/2}(y - y_i)) \end{aligned} \quad (6)$$

This estimator considers the bandwidth matrix as a function $H(y_i)$ of the sample points y_i . So different samples should have kernels with different sizes. $H(y_i)$ is then calculated as follows,

$$H(y_i) = h(y_i)I \quad (7)$$

where $h(y_i)$ is the distance from y_i to the k -th nearest point.

This method offers two advantages in calculating the bandwidth matrix. First, it avoids the under-smoothness and the over-smoothness problems. Second, it is an adaptive method and dependent on statistical data. When data are diverse and far apart, the kernel will be smoother. When data are tightly distributed, the kernel will be sharper. These are good properties for calculating the probabilities in Eq.(4), especially when the data size is small. Gaussian kernel is used in this paper and the density estimator in Eq.(6) becomes Eq.(8) and can be solved.

$$\hat{f}(y) = \frac{1}{(2\pi)^{q/2}N} \sum_{i=1}^N [(h(y_i))^{-q/2} \cdot \exp(-\frac{1}{2}(y - y_i)^T (h(y_i)^{-1}I)(y - y_i))] \quad (8)$$

The temporal saliency map estimated based on the above-mentioned method is effective to detect moving objects. However, when there is a serve changing of illumination within the sub-volume, the variance will also be large. To distinguish the difference between serve illumination changes and object motion, we would employ the Lambertian model [8] where a frame can be represented by the a product of illumination function $I(x, y)$ and reflection function $R(x, y)$, considering the temporal axis, this equation can be represented as follows:

$$f(x, y, t) = I(x, y, t) \cdot R(x, y, t) \quad (9)$$

If there is no motion information, the condition of Eq.(10) should be satisfied

$$R(x, y, t) = R(x, y) \quad (10)$$

In this case, reflection function will remain unchanged in the temporal sequence, then $f(x, y, t)$ is proportional to the illumination function $I(x, y, t)$. Since the spatio-temporal volume consists of a small number of frames (20 frames in our experiments), it is reasonable to assume that within a short period of time Δt (less one second in our experiments if we assume 30 frames per second), the following Eq.(11) are satisfied.

$$\begin{aligned} \lim_{\Delta x \rightarrow 0, \Delta t \rightarrow 0} \int_{\Omega_y} I(x_0 + \Delta x, y, t_0 + \Delta t) dy &= \int_{\Omega_y} I(x_0, y, t_0) dy \\ \lim_{\Delta y \rightarrow 0, \Delta t \rightarrow 0} \int_{\Omega_x} I(x, y_0 + \Delta y, t_0 + \Delta t) dx &= \int_{\Omega_x} I(x, y_0, t_0) dx \end{aligned} \quad (11)$$

where Ω_x represents any small neighborhood of x in a region where Eq.(10) is satisfied, and Ω_y represents any small neighborhood of y in a region where Eq.(10) is satisfied. Then Eq.(11) shows that if two pixels in 4-adjacency relation follow Eq.(10), their density functions in Δt will be similar. On the other hand, if a pixel does not follow Eq.(10), then $f(x, y, t)$ will not only dependent on illumination function $I(x, y, t)$, but also object reflection function $R(x, y, t)$. Their density function will be different. Therefore, to decide whether a particular pixel is under illumination, the Kullback-Leibler divergence between the distribution of this pixel and its 4 adjacent points are calculated. If the divergence is small, that pixel is classified as illuminated pixel without motion.

2.2.2 Computing Spatial Saliency

Computing the spatial saliency is similar with that in temporal saliency. Considering the density of x_0 in the spatial vector set: $X' = (x_0, x'_1, \dots, x'_{N'-1})$, the spatial saliency of x_0 can be computed using the following equation:

$$\begin{aligned} I_s(x_0) &= I_s(r, s) \\ &= -\log_2[P(Im_0(r, s)|B(r, s))] \\ &= -\log_2(P(x_0|X')) \end{aligned} \quad (12)$$

where $B(r, s)$ represents the set of spatial neighbor patches centering at $Im_0(r, s)$.

To reduce the computation time, before estimating the probability of $P(x_0|X')$ in Eq.(12), we employ cross-scale difference features [10] from local intensity, colors and orientations to verify whether x_0 is a spatially salient patch (SSP). This contrast feature based method has shown good performances in static image analysis.

A straightforward advantage of using spatial saliency is to detect slow motion. In the case an object slows down its speed, it loses its uniqueness in temporal domain, so its temporal saliency will decrease gradually, and when it is no longer a TSV, its temporal saliency becomes zero. To keep tracking the object saliency, we calculate the spatial saliency from all SSPs. As the spatial saliency will not be affected by object motion, the object will not lose its saliency even it is moving very slowly. An illustration of the object spatial saliency is shown in Figure 2(c).

2.2.3 Computing Spatiotemporal Saliency

After calculating the spatial saliency and temporal saliency of a patch, the next step is to fuse these two saliency maps. Ma et al. [13] generate the map by combining some arbitrary weighting factors to different components. Qiu et al [16] have shown that the spatiotemporal saliency of a video patch can be formulated as the conditional information with spatial and temporal contexts and under reasonable assumptions that the spatiotemporal saliency can be naturally expressed as the sum of the spatial saliency and the temporal saliency. Writing the self information of a patch in the spatial and temporal contexts we have

$$I(x_0) = -\log_2(P(x_0|X, X')) \quad (13)$$

That is, the information content of patch x_0 can be obtained from the minus logarithm of probability of x_0 given the conditions of both X and X' . This model can be simplified to Eq.(14), with the assumption that the spatial and temporal conditions are independent.

$$\begin{aligned}
I(x_0) &= -\log_2(P(x_0|X, X')) \\
&= -\log_2(P(x_0|X)P(x_0|X')) \\
&= -\log_2 P(x_0|X) - \log_2(P(x_0|X')) \\
&= I_t(x_0) + I_s(x_0)
\end{aligned} \tag{14}$$

As shown in Figure 1, x_0 can be either TSV or SSP. So there are four possible cases:

- Case 1: x_0 is TSV and SSP.

Since x_0 is both temporally salient and spatially salient, the spatiotemporal saliency of x_0 can be formulated as follows:

$$\begin{aligned}
I(x_0) &= I_s(x_0) + I_t(x_0) \\
&= -\log_2\left(\frac{1}{N'} \sum_{i=1}^{N'} K_H(x_0 - x'_i)\right) \\
&\quad - \log_2\left(\frac{1}{N} \sum_{i=1}^N K_H(x_0 - x_i)\right)
\end{aligned} \tag{15}$$

- Case 2: x_0 is a TSV but not SSP.

The spatiotemporal saliency of x_0 is the same as its temporal saliency:

$$I(x_0) = I_t(x_0) = -\log_2\left(\frac{1}{N'} \sum_{i=1}^{N'} K_H(x_0 - x'_i)\right) \tag{16}$$

- Case 3: x_0 is not a TSV but SSP.

The spatiotemporal saliency of x_0 is the same as its spatial saliency:

$$I(x_0) = I_s(x_0) = -\log_2\left(\frac{1}{N'} \sum_{i=1}^{N'} K_H(x_0 - x'_i)\right) \tag{17}$$

- Case 4: x_0 is not TSV nor SSP.

Since Both spatial and temporal saliency of x_0 are zero, the spatiotemporal saliency of x_0 is zero.

$$I(x_0) = 0 \tag{18}$$

Figure 2 illustrates how our spatiotemporal ISM can detect moving foreground objects when its states change from moving to stationary. Figure 2(a) shows three frames from the video "Browse2" from the CAVIAR dataset. When the person walks slowly and then stops, the temporal saliency of the person region becomes smaller as illustrated in Figure 2(b). When the person is not moving, the temporal ISM shows that area of the person has very low saliency. However, the spatial saliency would not be affected by the object motion speed as shown in Figure 2(c). It can be seen that the person does not lose his spatial saliency with different speeds. However, spatial ISM is noisy. To solve this problem, we consider locations that happen to be TSVs in their

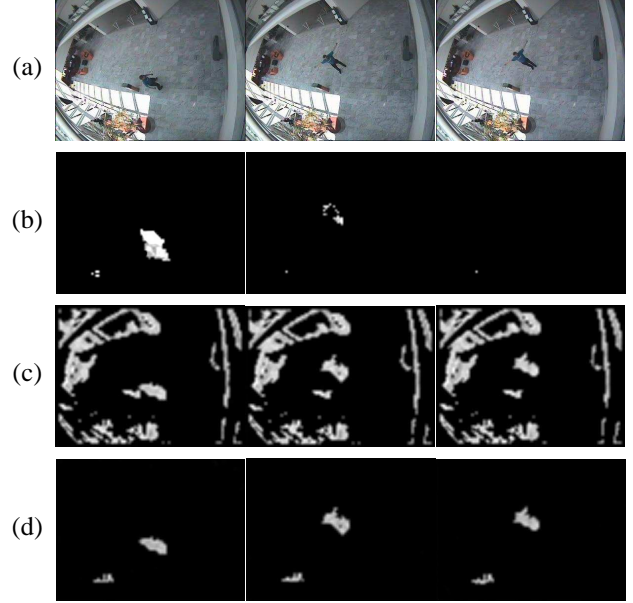


Figure 2. (a) Original frames, (b) Temporal ISM, (c) Spatial ISM, (d) Spatio-temporal ISM. The three columns are from Frame 70, 105, 140 respectively in CAVIAR "Browse2" database. This video clip contains two persons, one person walks to the center of the scene, keeps static for several seconds and walks away from the scene; the other person with slow motion keeps sitting at the bottom left corner.

temporal saliency map. Spatial saliency is used only when the object is slowing down or stop moving. Figure 2(d) shows the spatiotemporal ISM, which combines the spatial ISM and temporal ISM and it is shown that the areas covering the person show a high degree of saliency whether the person is moving or not.

3. Experimental Results

We have applied our spatiotemporal ISM to the detection of moving foreground objects in real video data. The experimental results are divided into two parts. First, we evaluate the performance of our proposed method using two datasets in CAVIAR [27], namely INRIA entrance hall and shopping mall front view. Second, the proposed method is compared with existing methods using CAVIAR INRIA entrance hall [27] and PETS2001 [26] datasets. Two methods, namely Mixture of Gaussians (MOG) [25] and Adaptive Kernel Density Estimation (AKDE) [14], are selected for comparison. In all experiments, we set $N = 20$ which is the number of frames in each sub-volume and $N' = 25$, the size of patch spatial surrounding region.

3.1. Evaluation of the proposed method

CAVIAR database [27] consists of 3 datasets, namely INRIA entrance hall, shopping mall front view and shop-

ping mall corridor view. The INRIA entrance hall dataset has six types of events, namely "Browsing", "Fighting", "Groups_meeting", "Leaving_bags", "Rest" and "Walking", totally 28 video sequences. These video sequences are captured from inclined look-down camera with a wide angle. The bottom left region of the video is under severe illumination condition. The shopping mall front view dataset consists of 26 video clips. This database is selected to evaluate our method because of significant illumination variations. Furthermore, people outside the shops have a shadow, which is challenging for the detection.

A typical experimental result from CAVIAR INRIA entrance hall is shown in Figure 3. Figure 3(a) shows the original video "Browse1" at frame 20, 30, 40 and 55, while the ISM and the detection results are shown in Figure 3(b) and (c) respectively. The rectangles show the detected foreground objects and the values indicate the information value of each region. The difficulty is to detect the three moving persons in the illuminated area. The lighting changes from time to time in this region, which makes the corresponding background very unstable. Moreover, when persons are passing this area, their appearance will change greatly because of the strong illumination reflection. The proposed spatio-temporal ISM gives good results, including the two persons with very slow motion on the left.

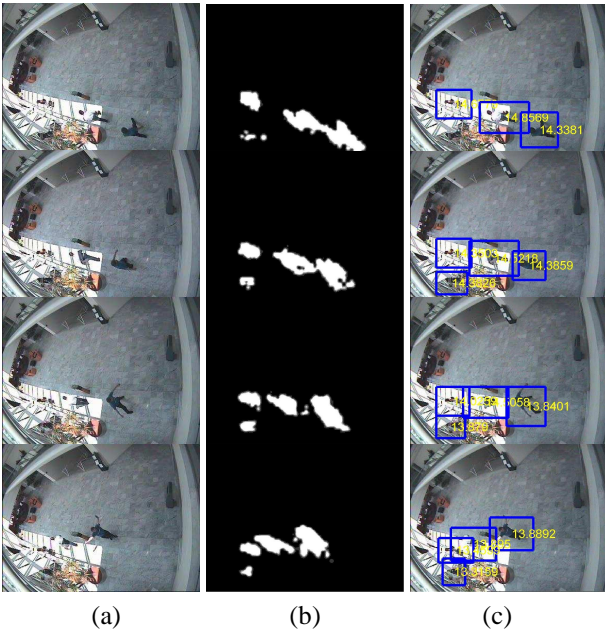


Figure 3. (a) Original video, (b) Spatio-temporal ISM, (c) Foreground detection rectangle result and object information content. From up to down rows:frame 20, 30, 40, 55 from video "Browse1" CAVIAR database. This video contains four persons moving with different speed, three of them are moving in the illuminated area. The rectangle value is obtained from averaging the spatio-temporal ISM where object motion is detected

Database	TP	FP	TG	FAR	DR
Browsing(6)	6924	193	7298	2.71%	94.88%
Fighting(4)	5214	197	5625	3.64%	92.69%
Groups_meet(6)	7550	172	7815	2.23%	96.61%
Leaving_bags(5)	7356	164	8702	2.18%	84.53%
Rest(4)	5005	152	5322	2.95%	94.04%
Walking(3)	5397	138	5512	2.49%	97.91%
ShopCenter(26)	46172	826	48753	1.76%	94.71%
Average	—	—	—	2.16%	93.92%

Table 1. Moving object detection results on CAVIAR database. TP:True Positive, FP:False Positive, TG:Total Ground truth, FAR:False Alarm Rate, FAR=FP/(TP+FP), DR:Detection Rate, DR=TP/TG. The number in the bracket represents the total number of video clips in each particular scenario

For the CAVIAR INRIA entrance hall and shopping mall front view datasets, ground truth data are available so that we can make quantitative analysis of our proposed method. The detection rate and the false detection rate of each video sequence are recorded and tabulated in Table 1. Our performance evaluation method is the same as the one in [1], where true positive (TP), false positive (FP), false negative (FN) and total groundtruth (TG) are used. The region is correctly detected if the overlapped area of the detected motion region and the ground truth bounding box is over ninety percent. Using this settings, the average detection rate (DR) using our proposed method is 93.92% while the average false alarm rate (FAR) is 2.16% as shown in Table 1.

3.2. Comparing the proposed method with existing methods

This section compares the proposed method with two existing methods, namely Mixture Of Gaussians [25] and Adaptive Kernel Density Estimation (AKDE) [14] using the INRIA entrance hall dataset in CAVIAR [27] and PETS2001 [26].

The improved adaptive Mixture of Gaussian model [25] is used to construct the probability density function for each pixel independently and pixel-level background subtraction is performed to find the regions of interest. Experimental results show that MOG is able to successfully model background that has regular variations. However, MOG does not perform well when there are severe illumination changes in a short period of time. Furthermore, when an object is with relatively small motion, MOG may mis-classify that region(s) as background and update the background accordingly. These situations can be illustrated using the example in Figure 4. This video clip from PETS2001 [26] is under fast illumination change when the sunlight is blocked by a piece of cloud. Another challenge in this video is that the tree is waving in the present of wind. From Figure 4(b), it can be seen that MOG cannot model the background well

rate on CAVIAR datasets are 93.92% and 2.16% respectively. Comparison with two popular methods, namely Mixture Of Gaussians and Adaptive Kernel Density Estimation, are also reported. Experimental results show that the proposed method is robust to illumination changes and no prior knowledge of the scene is required. Moreover, ISM not only provides the saliency of each pixel for object detection, but also gives additional higher level object information such as the object motion speed.

Our future work will be concentrated on exploring object saliency correlation between successive frames in a multi-dimensional space and make use of the ISM for event recognition.

References

- [1] F. Bashir and F. Porikli. Performance evaluation of object detection and tracking systems. *Proceedings 9th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, pages 7–14, 2006.
- [2] T. Brox, A. Bruhn, and J. Weickert. Variational motion segmentation with level sets. *European Conference on Computer Vision*, pages 471–483, 2006.
- [3] D. B. Bruce and J. K. Tsotsos. Saliency based on information maximization. *Advances in Neural Information Processing Systems*, pages 155–162, 2006.
- [4] A. Bugeau and P. Perez. Detection and segmentation of moving objects in highly dynamic scenes. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007.
- [5] A. Elgammal, D. Harwood, and L. Davis. Non-parametric model for background subtraction. *Proceedings of the 6th European Conference on Computer Vision*, pages 751–767, 2000.
- [6] W. Fang and K. L. Chan. Using statistical shape priors in geodesic active contours for robust object detection. *International Conference on Pattern Recognition*, pages 304–307, 2006.
- [7] V. Ferrari, T. Tuytelaars, and L. V. Gool. Object detection by contour segment networks. *European Conference on Computer Vision*, pages 14–28, 2006.
- [8] R. C. Gonzalez and R. E. Woods. Digital image processing(second edition). *Prentice Hall*, 2003.
- [9] M. Heikkila and M. Pietikainen. A texture-based method for modeling the background and detecting moving objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):657–662, 2006.
- [10] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.
- [11] T. Liu, J. Sun, N.-N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007.
- [12] Y. Liu, H. Yao, W. Gao, X. Chen, and D. Zhao. Nonparametric background generation. *International Conference on Pattern Recognition*, pages 916–919, 2006.
- [13] Y. F. Ma, L. Lu, H. J. Zhang, and M. J. Li. A user attention model for video summarization. *Proceedings of the tenth ACM international conference on Multimedia*, pages 533–542, 2002.
- [14] A. Mittal and N. Paragios. Motion-based background subtraction using adaptive kernel density estimation. *proceedings. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 302–309, 2004.
- [15] T. Parag, A. Elgammal, and A. Mittal. A framework for feature selection for background subtraction. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1916–1923, 2006.
- [16] G. Qiu, X. Gu, Z. Chen, Q. Chen, and C. Wang. An information theoretic model of spatiotemporal visual saliency. *International Conference on Multimedia & Expo*, pages 1806–1809, 2007.
- [17] S. P. N. Singh, P. J. Csonka, and K. J. Waldron. Optical flow aided motion estimation for legged locomotion. *IEEE International Conference on Intelligent Robots and Systems*, pages 1738–1743, 2006.
- [18] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. *proceedings. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1999.
- [19] A. A. Stocker. An improved 2d optical flow sensor for motion segmentation. *Proceedings of IEEE International Symposium on Circuits and Systems*, 2:332–335, 2002.
- [20] J. Sun, W. Zhang, X. Tang, and H. Shum. Background cut. *European Conference on Computer Vision*, pages 628–641, 2006.
- [21] Z. Yin and R. Collins. Belief propagation in a 3d spatiotemporal mrf for moving object detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007.
- [22] M. Yokoyama and T. Poggio. A contour-based moving object detection and tracking. *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 271–276, 2005.
- [23] T. Yu, C. Zhang, M. Cohen, Y. Rui, and Y. Wu. Monocular video foreground/background segmentation by tracking spatial-color gaussian mixture models. *IEEE Workshop on Motion and Video Computing*, 2007.
- [24] G. Zhang, J. Jia, W. Xiong, T.-T. Wong, P.-A. Heng, and H. Bao. Moving object extraction with a hand-held camera. *IEEE International Conference on Computer Vision*, 2007.
- [25] Z. Zivkovic. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recognition Letters*, 27(7):773–780, 2006.
- [26] <http://ftp.pets.rdg.ac.uk/pub/>.
- [27] <http://homepages.inf.ed.ac.uk/rbf/caviar/>.

Processing Spatial Queries in Wireless Sensor Networks

Yu Li

Abstract

In this paper we present our study on processing a special kind of spatial queries in wireless sensor networks. Wireless sensor networks have the potential to provide a wealth of information about the environments in which they are deployed. The sensor nodes are responsible for capturing environmental data and the base station is responsible for accepting and answering user queries. However, wireless sensor networks suffer from limited energy supply. Thus, energy efficiency is a key consideration in sensor network designs. We focus on a special kind of spatial queries returning areas under an average value constraint. We formulate the problem, study its hardness, and propose distributed solution for saving energy consumption in answering those queries.

1. Introduction

Wireless sensor networks have the potential to provide a wealth of information about the environments in which they are deployed. A wireless sensor network typically consists of a group of sensor nodes and a base station. The sensor nodes are responsible for capturing environmental data (e.g., temperature), and the base station is responsible for accepting and answering user queries (e.g., “Are there areas in the danger of fire disaster?”). Because the sensed data is stored on individual sensors, the user queries cannot be answered directly by the base station. In order to answer queries, collecting sensor readings is usually necessary. Otherwise, in-network processing is also a possible solution. Both of these approaches require communication between sensor nodes. However, a wireless sensor network is tempered by the limited energy supply of sensor nodes, and radio communication is the main energy consumer [1,6]. Therefore, it is of utmost importance to minimize the communication costs in order to improve energy efficiency of wireless sensor networks.

The queries on the wireless sensor network can be classified into several types focusing on different aspects. One of them, which has not been extensively discussed yet, is to query the spatial areas of the sensor network based on conditions. These conditions usually indicate some special events, such as being in danger of a fire

disaster. For example, “report the area of sensor nodes with average temperature higher than 50 °C”. In general, we may define it as a spatial query returning an area (in short, spatial query in this paper), which returns a set of sensors indicating an area. This query is useful in practice, such as in natural disaster monitoring system. In that kind of system we have huge wireless sensor network, but usually the result is just small areas. Collecting all sensor reading to the base station definitely answers the question, but may waste energy and short the life time of whole network. Is there any better strategy also giving the right answer but requiring less data collection? In this paper we try to answer this question on one special kind of spatial queries.

The special kind of spatial queries we studied is the query for areas under average requirement constraint. The example query previously given is an example. However, after modeling the wireless sensor network as a graph and the query as a combinatorial problem on it, we find that it is NP-hard. Achieving any polynomial optimal algorithm is even impossible with all sensor readings collected to the base station, and therefore we have to turn to approximate solutions, which is usually acceptable in practice. It is also not easy, because approximate algorithms in literature usually assume that all sensor readings are locally stored and accessible, which is equivalent to collecting them to the base station. However, we do find a technique which can identify sensor readings that has no chance to be in the result areas after studying those algorithms. With that technique, we set up an easier problem, namely budgeted shortest path problem, in the hope of saving as much energy as we can. We firstly establish a distributed algorithm based the classic Dijkstra’s shortest path algorithm. Then we extend it to more complex case to match the practice on current wireless sensor networks.

The rest of the paper will be organized as follows. Section 2 formulates the spatial queries we want to study, and study the challenges and present our basic idea. In Section 3, a complete distributed solution based on Dijkstra’s algorithm will be proposed as well as the extension. Then in Section 4, we briefly analysis the energy consumption in our algorithm, present it as the guide of applying our algorithm. Finally Section 5 lists the related work.

2. Problem Formulation

2.1 Description

Consider following query which query areas that may be in danger of fire disaster in a wireless sensor network monitoring temperature of a fortress.

```
Q1: SELECT area A
    FROM sensor_network S
    WHERE FOR ALL sensor s in A,
        Average(s.temp) > 50 ;
```

The ``Average(s.temp) > 50'' requires that the average temperature of the area must be higher than 50°C. The returning result is an *area*, which is a subset of the wireless sensor network consisting of sensor nodes. We define the area as

Definition area: An area in a wireless sensor network can only be (1) one sensor node, or (2) a set of sensor nodes (at least 2) which has at least one neighbor in the area.

In wireless sensor network, if a sensor node can directly communicate (without relay) with another one, it is called a neighbor of the other. So condition (2) ensures that the area is a connected graph.

However with this simple definition we may have many areas as candidates for specified average temperature constraint. For example, any subset of a candidate also fulfills the definition and constraint. Alternatively we want to find the biggest one, which is defined as

Definition biggest-area: Given an average threshold, an area in a wireless sensor network is biggest one when it maximizes the size while make the average value bigger than the threshold.

Notice that here we only consider the great-than predicate. However, other predicates, such as less-than, great-than-or-equal and less-than-or-equal, can be easily transformed. Therefore it is acceptable and in following discussion we focus on great-than predicate.

2.2 Problem Formulation

Wireless sensor network can be modeled as a undirected graph $G = \{V, E\}$, where each sensor node is a vertex in V . If two sensor nodes are neighbors, there is an edge in E connecting them. An arbitrary area is a connected sub-graph of G . Considering the result area of previous presented queries, we could limit it to be a spanning tree, because connected sub-graph can always transform into a spanning tree. The problem now is

transformed into finding the maximum spanning tree on G fulfilling the average constraint.

For specific threshold of average constraint, sensor nodes may sense different values, either being higher than the threshold or lower than it. For those who sensing a higher value (high node, for short), they will contribute to the area because they are able to probe other sensor nodes with lower value (low node, for short) into the area. High nodes are obviously in the future spanning tree we want, so the left work is to deliver their contribution as widely as possible. Based on this idea, we can further formulate our spanning tree problem as budgeted spanning tree problem [9].

In detail, we transform the problem as follows. For each vertex v_i of graph $G = \{V, E\}$ formulated from a wireless sensor network, denote t_i as the sensing value. Given a average threshold k , define the difference $\Delta_i = t_i - k$. Assume that there are some vertices $S = \{v_{s_1}, v_{s_2}, \dots, v_{s_m}\}$ with $\Delta_{s_i} > 0$ and denote $B = \sum \Delta_{s_i}$. Then our problem is to

maximize *size of T*

Subject to

T is a spanning tree in G

$S \subseteq T$

$\sum_{v_j \in S} |\Delta_j| < B$

Taking the sum of contribution of high nodes as the budget is reasonable because we can apply a two-phase process in practice. In first phase all high nodes report to base station their sensor readings, then in second phase base station can calculate the budget and notify the high nodes to perform an exploration process to deliver it. Low nodes affected by it will report itself to base station. In particular, the budget sometime can be calculated locally, sometime not. We will discuss the details in later sections.

2.3 Challenges

Unfortunately the budgeted spanning tree is proven to be NP-hard problem [9], which means that it is unlikely to find polynomial optimal algorithm even when we do not face the distributed environment. In other words, though naively collect all sensor readings to base station, optimally calculating the spanning tree is even impossible.

However we could turn to approximate result, which is usually still useful especially in application of temperature monitoring. Existing approximate algorithms for budgeted spanning tree are usually designed using prime-dual technique or Lagrangian relaxation technique, such as [9] and [8]. Unfortunately they all assume that the whole graph is stored locally, and highly rely on it

because of the iteration process in the calculation. As a research work focusing on application, we do not want to further improve that theoretical effort.

We observe that actually the approximate algorithms could also work with less sensor readings of lower nodes [9]. Those lower nodes are impossible to be included into final result areas if there is no path from any high node for passing contribution to it. Formally we say

Theorem 1: For any low node, if there is no path from any high node to it having the cost less than the budget, it is impossible to be included into any final result area. For a path, its cost is the sum of all nodes' differences.

The theorem is easy to prove: if such a node is possible to be included into a result area, the path along the spanning tree from any high node to it is must be such a path, which is a contradiction..

Having theorem 1 we may think about collect less sensor readings to save energy according to that. Intuitively we can filter sensor nodes by checking whether there is a shortest path from any high node with cost in the budget. This reminds us the Dijkstra's shortest path algorithm. However there are non-trivial challenges waiting us in distributed environment

1. A distributed Dijkstra shortest path algorithm should be adapted and designed. Furthermore, we are about to design a multi-source shortest path detecting algorithm.
2. Estimation on the possible message changing as well as energy consumption of the algorithm. If it is even bigger than directly collecting all readings of the network, we should abandon doing that. We need a mechanism to estimate.

In follow sections we will present our effort on design an efficient distributed Dijkstra shortest path algorithm suitable for our problem. After that, we will analysis the upper bound of energy consumption, which provides us a guide to apply this algorithm.

3. Distributed Solution

3.1 Single-Source Distributed Dijkstra's Algorithm

We start by adapting the Dijkstra's shortest path algorithm for the single-source problem, which is, we assume that there is only one high node v_0 in the graph. Formally, we want to
Find all vertices $v_i \in V (i \neq 0)$ which has the shortest path from v_0 to v_i with cost less than $B_0 = \Delta_0$.

Intuitively we can apply the Dijkstra's shortest path algorithm, which gradually doing breadth first search

starting from v_0 . The algorithm ensures that each vertex it discovers in each step has a shorter shortest path than other vertices not selected. So if we stop at the first vertex whose cost of shortest path is greater than B_0 , there will be no candidate vertices further. Otherwise, it should be discovered before the one we stopped. This is the basic idea, then we show how to do it in wireless sensor network.

Single-source Distributed Dijkstra's Shortest Path Algorithm

For each vertex v_i , maintain variables :

Report_flag = false

In_tree = false

Vertex v_0 , Initial Broadcast:

Calculate the budget $B_0 = \Delta_0$ (this will be carry with each Msg)

Prepare an array for v_0 's neighbors, namely neighbor_reply[], set all slots to be false

Initial a Priority-queue Q

If exist neighbor v_k with neighbor_reply[v_k] = false,

Broadcast with Msg = {explore, AC = 0 }

Else

algorithm end, report base station the Msg = {ending}

End if

Vertex v_i , On receive Msg = {explore, AC } from v_j :

If $AC + c_i \leq B_0$ and !In_tree,

reply v_j with Msg = {reply, AC + c_i }

Else If In_tree,

reply v_j with Msg = {In_tree}

Else If !In_tree and $AC + c_i > B$,

reply v_j with Msg = {Stop}

End if

Vertex v_i , On receive Msg = {In_tree} or Msg = {Stop} from v_j :

Mark neighbor_reply[v_j] = true

Vertex v_i , On receive Msg = {reply, AC } from v_j :

If v_j 's AC is smaller than the one in Q ,


```

replace it with new  $AC$ 
Else if  $v_j$  is not in  $Q$ ,
    Store  $v_j$  with key= $AC$  in  $Q$ ,
End if
If all neighbors have replied (checking array neighbor_reply),
    If  $Q$  is empty,
        algorithm end, report base station the Msg = {ending}
    Else
        Extract  $v_k$  with minimal  $AC$  in  $Q$ 
        Send  $v_k$  with Msg = {next_round,  $Q$ }
    End if
End if

```

Vertex v_i , On receive Msg = { next_round, Q } from v_j :

```

Report itself to base station with Msg = {report,  $t_i$ }
Prepare an array for neighbors, namely neighbor_reply[], set all
slots to be false
If  $v_j$  is in neighborhood, set neighbor_reply[ $v_j$ ] = true
If exist neighbor  $v_k$  with neighbor_reply[ $v_k$ ] = false,
    Broadcast with Msg = {explore,  $AC$ }
Else
    If  $Q$  is empty,
        algorithm end, report base station the Msg = {ending}
    Else
        Extract  $v_m$  with minimal  $AC$  in  $Q$ 
        Send  $v_m$  with Msg = {next_round,  $Q$ }
    End if
End if

```

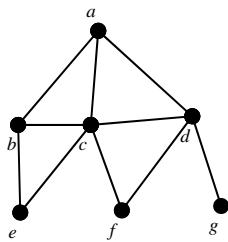


Figure 1 Example of Single-source Dijkstra's Algorithm. a is the start vertex with budget 8; at the exploration interval, sensor readings are: $b=1, c=2, d=3, e=5, f=4, g=6$.

To understand how the algorithm works, let us check the example in Figure 3. The detail of execution is shown in Table 1. Looking at step 1, vertex a broadcasts with the

initial value 0 (as well as the budget 6 and each step later), then wait for reply of b, c and d . After receiving the reply messages, it updates the empty priority queue Q , and select the least one as the next round start point, which is b . Followly in step 2 b receives the explore instruction as well as Q , and does the same work. However, this time its neighbors will reply the accumulated cost based on b 's cost, which is $1+2=3$ for vertex c , and $2+4=6$ for vertex e . Once again the least node c in Q is selected and the algorithm keeps on going until the Q becomes empty. Finally a extends 5 low nodes in 5steps.

Table 1 The execution detail on Figure 1. (#=in_tree, S=stop)

Step	Q	Action	Reply	Next
1	{}	a bcast 0	{b=1,c=2,d=3}	b
2	{c=2,d=3}	b bcast 1	{c=3,e=6}	c
3	{d=3, e=6}	c bcast 2	{a=#,d=5,e=7,f=6}	d
4	{e=6,f=6}	d bcast 3	{a=#,f=7,g=S}	e
5	{f=6}	e bcast 6	{b=#,c=#}	f
6	{}	End		

3.2 Extending to Multi-Source Case

Conventionally extending single-source Dijkstra's algorithm to multi-source case employs technique adding virtual vertex. By introducing a vertex connecting to all sources and setting all sources' cost to be zero, we can apply almost the same algorithm to find shortest path for low nodes. However it does not consider the budget constraint in that kind of simple extending. This is different between our problem and the conventional case.

The budget constraint in our problem makes the extending non-trivial. Consider high nodes in Figure 2. If two high nodes are adjacent (shown in Figure 2(a)), which can directly reach each other, we can apply the same technique of introducing virtual vertex. In particular we even do not need to do that thing. Instead by setting high node's cost to be zero, starting from v_i we will firstly extend to v_j before any low node. Thus the budget could be simply calculated by $B_{ij} = \Delta_i + \Delta_j$ when we start next round exploration at v_j . Considering even more high nodes, as long as they are adjacent to each other, we could apply same technique and similarly calculate the budget.

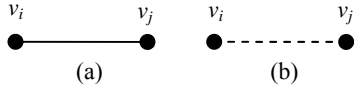


Figure 2 (a) adjacent high nodes, (b) indirectly reachable high nodes

On the other hand, simply introducing virtual vertex to connect indirectly reachable high nodes does not work. The problem arises in two aspects. On one hand we can not easily calculate new budget by sum. Assume that we connect two indirectly reachable high nodes such as v_i and v_j (shown in Figure 2(b)), and calculate the new budget as $B_{ij} = \Delta_i + \Delta_j$. When extend to some low node such as v_k , we actually assume that the path $v_i \mapsto v_j \mapsto v_k$ or $v_j \mapsto v_i \mapsto v_k$ (here \mapsto indicates indirectly connection) with cost smaller than B_{ij} , which further implies that the shortest path $v_i \mapsto v_j$ or $v_j \mapsto v_i$ has a cost smaller than Δ_i or Δ_j . This is not always expectable when we apply such a distributed algorithm. On the other hand, even we can ensure previous condition, we also may face some vertex v'_k which can only be included in the way such as $v_i \mapsto v'_k$ & $v_j \mapsto v'_k$ (v'_k accepts both contribution from v_i and v_j). Obviously, this is not a simple shortest path, so that it can not achieved by simply applying Dijkstra's algorithm.

Our basic idea to overcome previous difficulties is to employ a multi-phase exploration process to handle indirectly reachable high nodes, while keep the simple merging technique for adjacent high nodes the same. In follows we present our effort on how to extend previous single-source Dijkstra's algorithm.

3.2.1 Before First Batch of Exploration

Before first batch of exploration, each high node will report the base station itself as a candidate. We can utilize this to merge adjacent vertices. The idea is to run a checking process grouping reported high nodes according to whether they are directly reachable to each other. Then select each group a vertex as the start point of the exploration process in future.

Furthermore this checking can be designed asynchronously. Ideally our strategy should only allow one vertex as the exploration start point for each group. If we have an oracle telling us after what time there will be no high node reporting, achieving that goal is easy. Though theoretically that oracle can be implemented as waiting for sufficient long time, the practice prefer to an asynchronous strategy. The problem is that asynchronous

strategy may make one group having more than one start point. For example, consider the group $v_1 \leftrightarrow v_2 \leftrightarrow v_3$. If they arrive in the order v_1, v_3, v_2 , asynchronous strategy may select v_1 and v_3 as start point after first two arriving. In later exploration, v_2 will be selected by both exploration process from v_1 and v_3 , and it causes ambiguity problem. With the merge process which will be later introduced, it can be avoided by simply applying FIFO message handling on v_2 which accepts the first exploration. Thus we give the algorithm on base station as below

Base Station, On receive Msg = {report, t_j } from vertex v_j :

If $t_j > k$ then

If there is group S_i having vertex adjacent to v_j ,

$$S_i = S_i \cup \{v_j\}$$

Send Msg = {Wait} to v_j

Else

Create new $S' = \{v_j\}$, store in base station

Send Msg = {Init_explore} to v_j

End if

End if

For vertices, on receiving wait message from base station, it waits for further exploration message; on receiving init_explore message, it starts initial exploration immediately. When explore another high node, budget is recalculated together as $B_{ij} = \Delta_i + \Delta_j$.

3.2.2 Merge Registration and Elimination

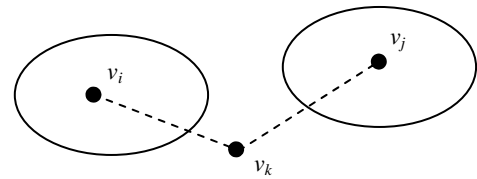


Figure 3 Before merge action taking place

In this section we introduce the merge mechanism to handle indirectly reachable high nodes. Checking the case in Figure 3, we have two exploration processes started from v_i and v_j . There is another vertex v_k (suggest it is a low node now, we will consider the high node case later). It may be explored by v_i or v_j . W.o.l.g., suggest it is firstly explored by v_i . Later when it is also explored by v_j , we know that those two exploration processes should be merged, because with the help of vertex v_k , the

contribution of v_i and v_j could be combined together to probe even smaller low node.

We can stop the exploration process of v_j immediately as it just arrives v_k , but the exploration process of v_i is still on-going and we do not know where it arrives. We make v_k send base station a message $\text{Msg}=\{\text{merge}, v_i, v_j\}$. On receiving that message, base station registers an entry for merging the exploration process of v_i and v_j later, and wait for the end message of v_i arriving. On receiving that message, base station eliminates the merge registration and sends new init_explore message to v_i or v_j with new budget $B_{ij} = \Delta_i + \Delta_j$. On receiving that, v_i or v_j firstly spreads the message in the tree of last exploration results from v_i and v_j , which is now connected by a internal node v_k . After that we can decide the new selected node and continue the Dijkstra's algorithm in next round.

Theorem 2: The result of merged exploration process is at least equal to the union of v_i and v_j .

Proof: For any low node can be explored, such as v_n .

1. If it has a shortest path from v_i with cost in Δ_i , it has been explored before merge action taking place;
2. If it has a shortest path from v_j with cost in Δ_j , it could be either explored or not before merge action. If its shortest path even shorter than v_k 's, it is explored before merge action. Otherwise, its shortest path's cost is also smaller than $B_{ij} = \Delta_i + \Delta_j$, so it must be explored sometime later.

Considering that there may be some low node being explored by the combinatorial contribution, the theorem follows.

Remember that v_k can also be a high node. If it does not equal to v_j itself or not directly connected with v_j , it actually represents another exploration tree, so we should merge with that one first. Otherwise we can also apply same merge registration technique. Marking each node's owner exploration tree is necessary here, but it is easy to implement while selecting nodes by assigning the start point as the identification.

Merge technique can also overcome the multi start point problem left in last section. As v_1 and v_3 are selected to be start point of the exploration process. Next, v_2 should be explored by only v_1 or v_3 because of the FIFO

strategy. After that, either v_1 's message reaching v_2 , or v_3 's message reaching v_1 through v_2 , the merge registration will take place. Know from theorem 2, we will not miss anything because we have late-merge action.

The end checking mechanism in the base station also needs change. Because we introduce merge mechanism, the algorithm will end only when all end messages of exploration process are received, and there is no more merge entries to be eliminated.

4. Energy Consumption Estimation

Intuitively, the distributed solution will save some energy consumption by filtering unnecessary report to base station. That is why we design such a distributed computing mechanism. However in order to maintain the distributed mechanism, additional overhead is unavoidable. Whether to apply the distributed way or just simply collecting all sensor readings depends on how good result we can achieve. Therefore estimation of energy consumption is not only necessary, but also critical to make the mechanism complete.

Estimation of Single-source Distributed Dijkstra Algorithm

Suggest that the final result set is A . Known from the algorithm, it is a sub graph (precisely, a sub spanning tree) of whole wireless sensor network. Denote it as $A = \{V_A, E_A\}$, where V_A is the set of vertices and E_A contains all edges between them.

First, there will be report cost for each vertex. Denote the sum of report cost on A as $R(A)$. Then consider the message changing cost in exploration steps. To generate the result, each vertex only explores its neighbors by broadcasting once. Therefore the cost for broadcasting is $|V_A| \cdot M_b$, where M_b is the cost for single time broadcasting in exploration purpose. And then neighbors reply the explorer with directly sending messages. This kind of communication will happen on each edge in E_A for two times. Vertex on either side of an edge will invoke a round of message communication. In all the cost is $2|E_A| \cdot M_e$, where M_e is the cost for single round of sending and receiving. Remember that the vertices not to be the candidate but in the neighborhood of A also reply the exploration at least once with stop message. If denote the neighborhood of A as $ne(A)$, the total cost should be $|ne(A)| \cdot M_e$. Finally we have the summary of the cost in single-source distributed Dijkstra algorithm as

$$R(A) + |V_A| \cdot M_b + (2|E_A| + |ne(A)|) \cdot M_e$$

Estimation of Multi-source Distributed Extension

Multi-source case is more complex when it involves merge registration and elimination process. Its recursive characteristic makes the estimation difficult. In general,

when two sub areas such as A_1 and A_2 of the final result A_f need to merge, a full scanning of the tree of A_1 and A_2 is unavoidable, which causes additional cost $(|E_{A_1}| + |E_{A_2}|) \cdot M_e$. The difficult is hard to predict how many times the merge action will take place, though it has an upper bound which equals to the number of high nodes in A_f . Therefore for general case we can only estimate it according to the upper bound, which is

$$R(A_f) + |V_{A_f}| \cdot M_b + (2|E_{A_f}| + |ne(A_f)|) \cdot M_e + |E_{A_f}| \cdot |hi(A_f)| \quad (1)$$

where $hi(A_f)$ is the set of initial high nodes in A_f .

Cost of Transmitting Q

The priority queue Q used in the distributed algorithm may cause extra energy consumption. When it exceeds the packet size of single time communication, in order to fully transmitting it multiple messages are necessary. It is easy to see that the bigger the result area is, the easier Q causes that problem. Therefore sometime we have to consider the overhead bought by transmitting Q . However, it is also not easy because Q is changed along with every different exploration process, and we are still in developing a model for it.

Application of Estimation Results

Estimation could help us to decide whether to apply the distributed algorithm or not. For the result area, we present the reporting cost as $R(A_f)$. Similarly the cost of collecting the whole graph is $R(G)$. They are both predictable as base station usually having the topology of the network. By comparing $R(G)$ to the estimated cost of distributed algorithm, we decide whether go on in distributed way, or just switch to naïve collecting.

However we still assume that we know final result before estimation, which seems to be a barrier in practice. Actually, this could be eliminated by reusing last time result or based on the statistic analysis, because when monitoring in practical system, the changes usually will not change suddenly.

5. Related Work

The research in wireless sensor network is active in recent years [1-4]. Processing spatial queries is one of the important sub areas. Different to previous research in distributed computing, which focuses on improving execution time, nowadays research emphasizes saving energy to extend network life, such as works of [5], [7] and [13].

The model of our query, budgeted spanning tree problem, is studied extensively in past literature. It is

firstly introduced by Guha, et. al. [9] and proved to be NP-hard. Approximate algorithms for it is proposed in [9] and later in [8] using Lagrangian relaxation technique. This problem is also studied under the name of minimal spanning tree subject to a side constraint in [10] and [14].

The easier problem, distributed shortest path problem was studied in [11] and [12]. The best result present in [11] for this problem, which is also based on Dijkstra's algorithm is $O(E^{1+\epsilon} \cdot \log V)$ in message complexity and $O(V^{1+\epsilon} \cdot \log V)$ in time complexity. However, we can not apply that one because it assumes the whole network is accessible and does not have a budget constraint. Instead we adopt the idea of merging and design our own distributed algorithms in consideration of budget constraint and energy consumption.

6. References

- [1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks," IEEE Communications Magazine, vol. 40, no. 8, pp. 102–114, 2002.
- [2] S.R. Maden, M.J. Franklin, J.M. Hellerstein, and W. Hong, "TinyDB: An Acquisitional Query Processing System for Sensor Networks," ACM TODS, vol. 30, no. 1, pp. 122–173, 2005.
- [3] D.J. Abadi, S. Madden, and W. Lindner, "REED: Robust, Efficient Filtering and Event Detection in Sensor Networks," Proc. of VLDB'05, pp. 769–780, 2005.
- [4] U. Srivastava, K. Munagala and J. Widom, "Operator Placement for In-Network Stream Query Processing," Proc. of PODS'05, pp. 250–258, 2005.
- [5] A. Silberstein, R. Braynard, and J. Yang, "Constraint Chaining: On Energy-Efficient Continuous Monitoring in Sensor Networks", Proc. of SIGMOD'06, pp. 157–168, 2006.
- [6] T. S. Rappaport, "Wireless Communications: Principles and Practice," Prentice Hall, 1996.
- [7] X. Tang and J. Xu, "Extending Network Lifetime for Precision-Constrained Data Aggregation in Wireless Sensor Networks," Proc. of Infocom'06, pp. 1–12, 2006.
- [8] A. Moss and Y. Rabani, "Approximation algorithms for constrained for constrained node weighted steiner tree problems," Proc. Of STOC'01, pp. 373–382, 2001.
- [9] S. Guha, A.Moss, J.Naor and B. Schieber, "Efficient Recovery from Power Outage (extended abstract) ," Proc. Of STOC'99, pp. 574–582, 1999.
- [10] V. Aggarwal, Y. P. Aneja and K. P. K. Nair, "Minimal spanning tree subject to a side constraint, " Computers & Operations Research, Vol.9, No.2, pp. 287–296, 1982.
- [11] B. Awerbuch, "Distributed Shortest Paths Algorithms (Extended Abstract), " Proc. Of STOC'89, pp. 490–500, 1989.

- [12] B. Awerbuch, "Shortest Paths and Loop-Free Routing in Dynamic Networks, " Proc. Of SIGCOMM'90, pp. 177-187, 1990.
- [13] M.L. Yiu, N. Mamoulis and S. Bakiras, "Evaluation of Spatial Pattern Queries in Sensor NetWorks," HKU CS Tech Report TR-2007-02, 2007.
- [14] S. Pirkwieser, G. R. Raidl and J. Puchinger, "A Lagrangian decomposition/evolutionary algorithm hybrid for the knapsack constrained maximum spanning tree problem, " Technical Report TR 186-1-07-03, Institute of Computer Graphics and Algorithms, Vienna University of Technology, 2007.

Data Management on Flash Storage

ON Saitung

Abstract

As high-density flash memory has been widely adopted as data storage media, issues on how to provide efficient data management on flash memory have started receiving growing attention in recent years. After a brief introduction on the characteristics of flash memory, this paper reviews some typical flash memory technologies. Through detailed discussion on the limitation of the current flash memory technology, some possible directions for future research are drawn out.

1. Introduction

With the recent technology breakthroughs in both capacity and reliability, flash memory has been increasingly adopted as data storage media for a wide spectrum of computing devices. Like magnetic disk drives, flash memory is non-volatile and retains its contents even when the power is turned off. Compared with magnetic disk drives, flash memory has its superiority such as lighter weight, smaller size, better shock resistance, lower power consumption, less noise, and faster read performance [1]. As the capacity increases and price drops, flash memory will compete more successfully with lower-end, lower-capacity magnetic disk drives [2]. This trend arouses researchers to develop new flash-based storage system.

Although flash memory has aforementioned advantages, it has many disgusting nature features. These unfavorable features include: (1) No In-Place Update, and (2) The endurance issue. These features bring two major issues for the flash memory storage system implementation: (1) Garbage collection policy, and (2) Wear-leveilling mechanism.

In the past work, various techniques were proposed to overcome these limitations of flash memory, and to exploit unique characteristics of flash memory to achieve best attainable performance for flash-based storage system.

The rest of this paper is organized as follows. Section 2 discusses the characteristics of flash memory, and introduces the issues brought by these unique features. Section 3 reviews some typical flash memory technologies. In section 4, we analyze the limitation of the existing techniques, and elicit some possible research directions. Lastly, section 5 summarizes the content of this paper.

Table 1. Access Speed: Magnetic disk vs. NAND Flash

Media	Access time		
	Read	Write	Erase
Magnetic Disk	12.7 ms (2 KB)	13.7 ms (2 KB)	N/A
NAND Flash	80 μ s (2 KB)	200 μ s (2 KB)	1.5 ms (128KB)

2. Features of Flash Memory

In this section, we describe the key features of flash memory which distinguish itself from magnetic disk drives. Furthermore, we discuss on how they would affect the design of the flash-based storage system.

2.1. The Structure of Flash Memory

There are two major architectures in flash memory design: NOR flash and NAND flash. NOR flash is a kind of EEPROM, and NAND flash is designed for data storage. In this paper, we will focus on NAND flash. Hereafter, we use the term flash memory to refer to NAND-type flash memory in this paper.

A NAND flash memory chip is organized in many blocks and each block is of a fixed number of pages. A block is the smallest unit for erase operation, while reads and writes are handled by pages. The typical block size and page size of flash memory is 16KB and 512 bytes, respectively.

2.2. Asymmetric Speed of Read/Write

Unlike magnetic disk drives whose read and write speed are the same, flash memory has asymmetric read and write speed. As flash memory is a pure electronic device, it takes longer to write a cell until reaching a stable status than to read the status from a cell. As is shown in Table 1, its read speed is at least twice faster than write speed. We can also notice that flash memory has much faster speed than magnetic disk on both read and write operations. Upon observation, it is significant for a flash-based storage system to reduce write operations, even at the cost of increment of read operations, as long as the overall performance enhances.

2.3. No In-Place Update

With flash memory, the existing data items cannot be overwritten directly. Instead, in order to update a data item in place, a time-consuming erase operation must be performed before overwriting. What makes it worse is that the erase operation cannot be performed selectively on the particular data item or page, but can only be done for an entire block of flash memory called erase unit containing the data. Updating data in place is not efficient because all data in the block must first be copied to a system buffer, and then updated. Then, after the block has been erased, all data must be written back from system buffer to the block. Thus, updating even one byte requires one erase and several write operations. The update performance of the flash-based storage system may degrade greatly.

One approach which addresses this issue is out-place-update scheme. Under this scheme, for each update, the new version of the data will be written to some available space elsewhere. The old version of the data is then invalidated and considered as “dead”. The latest version of the data is considered as “live”. This scheme successfully reduces the requests of an erase operation. However, this scheme brings new problem that we need to recycle the space occupied by the dead data when the available space is low. That is so called garbage collection. A well-designed garbage collection policy should try to minimize its overhead, due to live data copies.

Another approach is in-page-logging scheme. The basic concept of this scheme is: whenever an update operation is to be performed, instead of overwriting the existing data, only the changes are appended. This scheme efficiently avoids frequent erase and write operations, at the cost of increasing read operations.

These two approaches will be further discussed in section 3.

2.4. Limited Erase Cycle

In flash memory, each block has a limit on the number of erase cycles (typically up to 100,000 times). A worn-out block will suffer from write errors. In order to lengthen the life span of flash memory, a mechanism called wear-levelling should be adopted to ensure that erase cycles are evenly distributed across the entire segment of flash memory.

According to this feature, it is essential to find an efficient wear-levelling mechanism to prolong the life span of flash memory.

2.5. Uniform Access Speed

Flash memory is a purely electronic device and thus has no mechanical latency when accessing data. Therefore, flash memory can provide uniform random

access speed. Unlike magnetic disks whose random access speed is much slower, the time to access data in flash memory is nearly linearly proportional to the amount of data regardless of their physical locations. This property is among the key features which can be taken advantage of to enhance the performance of the flash-based storage system.

3. Typical Flash Memory Technologies

In the past, many researches have been done on the issues of flash memory management. Excellent research results and implementations were reported on performance enhancement, especially on garbage collection and system architecture designs. Due to space limitations, we select some typical technologies to review.

3.1. Flash Translation Layer

Kawaguchi, et al. [3] proposed a flash memory translation layer to provide a transparent way to access flash memory through the emulating of a block device. With FTL, flash memory appears to upper layers like a disk drive, application algorithms and access methods will function adequately without any modifications. The objective of FTL is to encourage a quick deployment of flash-memory technology.

However, due to the aforementioned limitations of flash memory, this approach is not likely to yield the best attainable performance. Under this scheme, to update a data item, it must be preceded by erasing the erase unit containing this data. Observed from table 1, an erase operation takes much longer time than read/write operations. Also, as it does not consider the endurance issue, this approach may shorten the life span of the flash memory. Therefore, some high-level management algorithms, such as dynamic physical/logical address translation, garbage collection, and wear-levelling are necessary to be implemented upon FTL.

3.2. Garbage collection

Just as mentioned in Section 2.3, out-place-update scheme successfully reduces the requests of an erase operation. Meanwhile, as a side-effect, garbage collection could be initiated when flash-memory storage systems have a large number of live and dead data mixed together. Garbage collection could have overhead due to the live data copies. Garbage collection policy should solve the following problems:

- a. When should the garbage collection execute?
- b. Which blocks should be erased during garbage collection?

- c. Where to write the live data?
- d. How to minimize the overhead due to the live data copies?

Researchers have proposed excellent garbage-collection policies. Kawaguchi, et al. [3] proposed a cost-benefit policy with a value-driven heuristic function for block recycling. Chiang, et al. [4] refined the above work and proposed a Cost Age Times (CAT) policy to guide the block recycling by considering the locality in the runtime access patterns. Chang and Kuo [5] introduced a real-time garbage collection mechanism to provide QoS guarantees for performance-sensitive applications.

There are mainly three policies which are adopted to decide which blocks to be recycled. The greedy policy selects a block containing the largest number of dead pages; the cost-benefit policy chooses the most valuable block according to the formula:

$$\frac{\text{benefit}}{\text{cost}} = \frac{\text{age} \times (1 - u)}{2u}$$

where u is the percentage of valid data in the block to be erased and age means the elapsed time since the block was created; the Cost Age Times policy chooses the block minimize the formula:

$$\text{CleaningCost}_{FlashMemory} * \frac{1}{\text{age}} * \text{NumberOfCleaning}$$

The cleaning cost is defined as the cost of every useful write: $u/(1-u)$. Cleaning times means the numbers of erase operations conducted on blocks.

The cost-benefit policy is more reasonable than the greedy policy because it takes the cost of erasing into account and give blocks just erased more time to accumulate garbage for reclamation. The CAT policy further considers giving blocks with fewer erase times more chances to be selected for erasing. This avoids concentrating garbage collection activities on a few blocks, thus allowing more even wearing.

Nevertheless, none of them gives a good solution to question d. It is an intractable but interesting topic for us to explore.

3.3. Wear-levelling

Just as mentioned in Section 2.4, the number of erase operations for each block is limited. A wear-levelling mechanism need to be implemented to ensure that erase cycles are evenly distributed across the entire segment of flash memory. Various methods have been proposed on this issue. Kim and Lee [6] proposed to periodically move live data among blocks so that blocks have more even life-times. Approaches were proposed to distribute hot data over flash memory for wear levelling in [3, 4]. Wear-levelling mechanism restricts the erase operations among blocks to prolong the life span of flash memory.

3.4. Physical/logical address translation table

For flash memory management, data are moved over flash memory from time to time, due to out-place updates, garbage collection, and wear-levelling. To resolve the residing location problem for data on flash memory, the concept of physical/logical address translation table is adopted, and each entry of the table (indexed by the logical block address, so-called LBA) contains the physical address of its corresponding LBA. Whenever the data of a block are moved to another block, this table should be updated accordingly.

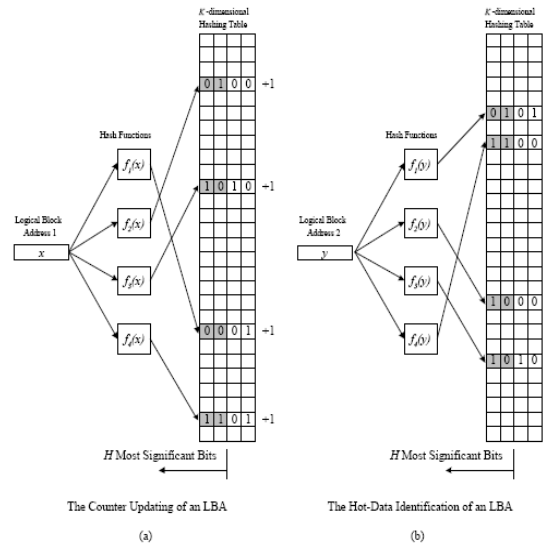


Figure 1: The counter updating and the hot-data identification of an LBA, where $C=4$, $K=4$, and $H=2$.

3.5. Hot data identification

Many researchers [3, 4] point out that on-line access patterns would have a strong impact on the performance of flash-based storage systems. Many methods have been proposed in the identification of hot and cold data. One trivial method on this issue is through tracking of data access time. Apparently, this method introduces significant memory foot-print. Hsieh, et al. [7] propose a highly efficient method for online hot-data identification with scalability considerations on precision and memory overheads. Within this method, a multi-hash-function framework is proposed to reduce the chance of false identification of hot data and provide excellent performance for hot-data identification. Here, detail about this method is given below:

As shown in figure 1, K independent hash functions are adopted to hash a given LBA into multiple entries of a M -entry hash table to track the write number of the LBA,

where each entry is associated with a counter of C bits. Whenever a write is issued to the FTL, the corresponding LBA is hashed simultaneously by K given hash functions. Among the corresponding K counters, the one with minimum counter value is incremented by one to reflect that the LBA is written again. When a counter reaches its maximum value, it is left unchanged. For every given number of sectors have been written, the values of all counters are divided by 2 in terms of a right shifting of their bits. This is an aging mechanism to exponentially decay the values of all write numbers as time goes on. Whenever an LBA is to be verified as a location for hot data, the LBA is also hashed simultaneously by the K hash functions. The LBA is judged to have hot data if the H most significant bits of every counter of the K hashed values contain a non-zero bit value. Because hashing tends to map a large address space into a smaller one, the memory overhead is low. The objective behind the adopting of K independent hash functions is to reduce the chance for the false identification of hot data. This approach gives an elegant solution for on-line hot data identification with very limited memory space.

3.6. In-Page-Logging approach

Lee and Moon [8] propose a new design called In-Page-Logging approach for flash memory based database servers. As aforementioned in Section 2.3, the basic idea of the In-Page-Logging scheme is: whenever a page is updated, only the changes made to this page are written to the database on the per-page basis, instead of writing the page in its entirety. Additionally, the data page and its log records are stored in the same erase unit. When there is no enough space for writing logs, a merge operation which applies the log records to their associated data page is performed to recycle space.

As illustrated in figure 2, the recent access data pages and their logs will be buffered in memory, and each physical block is divided into two segments - one for data pages and the other for log sectors.

Whenever an update is performed on a data page, the in-memory copy of the data page is updated. In addition, a physiological log record on the per-page basis is added to the in-memory log sector. When the in-memory log sector becomes full or when a dirty data page is evicted from the buffer pool, the associated in-memory log records are flushed to flash memory. When a block runs out of free log sectors, the data pages and log sectors are merged into a new block and thus recycle space. When a data page is to be read from flash memory due to a page fault, the current version of the page has to be computed on the fly by applying its log records to the previous version of the data page.

With this design, consequently frequent erase operations can be avoided. Furthermore, for each update,

the previous version of the data page remains intact in flash memory, but is just augmented with the update log records. For read operations, there are additional overhead for both IO cost (to fetch log records from flash memory) and computational cost (to compute the current version of a data page). As the benefits brought by the decrease of write and erase operations outweigh the additional cost of read operations, the overall performance is eventually improved.

The basic IPL design can be easily augmented to support transactional database recovery. As discussed before, an in-memory log sector is required to flush to flash memory when it becomes full or its associated data page is evicted from the buffer pool. In addition to that, in order to support transactional recovery, the IPL buffer manager has to force out an in-memory log sector to flash memory, if it contains at least one log record of a committing transaction. When there is no enough space in a block, the data pages and their committed log records are merged into a new block, together with their log records in active status. To read a data page from flash memory due to a page fault, the current version of the page is computed on the fly by applying its committed log records to the previous version of the data page. Through these new operation logics, the IPL design can efficiently support transactional database recovery.

The IPL approach gives an elegant solution for designing efficient flash-based database applications. However, this approach does not take the access pattern into account. It assigns the same size of log region for every block, regardless how many hot data pages are located. Apparently, “hot” blocks (contains more hot data) should have larger log region in order to further lower the frequency of erase operations. This is another interesting topic for us to explore.

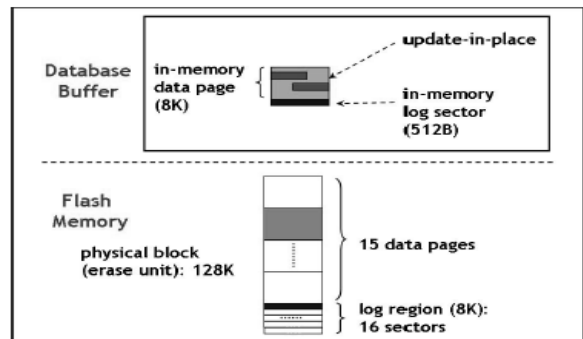


Figure 2: The design of In-Page-Logging

4. Future Research Directions

Although researchers have conducted many excellent researches on flash memory, there are still many

interesting topics to explore. We decide to dedicate ourselves to the following topics:

a. Designing new garbage collection algorithm to minimize the overhead due to the live data copies.

In real-world applications, updating a data page is likely to be quickly followed by updates to the same or related data pages (Known as update locality). Therefore, the overhead could be possibly reduced if the garbage collection algorithm smartly adjusts the locations of data pages, allowing data pages which are updated almost at the same time to reside in the same block.

b. Refining the In-Page-Logging approach to further improve the overall performance of flash-based database applications.

Upon observation, “hot” blocks should have fewer data pages and larger log region, while “cold” blocks should have more data pages and smaller log region. Our objective is to design an on-line algorithm to maintain such mechanism.

5. Conclusion

In addition to the flash characteristics, this paper reviews the related work on flash memory. Through detailed discussion on the limitation of the current flash memory technology, some possible directions for future research are drawn out.

6. References

- [1] Fred Douglass, Ramon Caceres, Frans Kaashoek, Kai Li, Brian Marsh, and Joshua A. Tauber, Storage Alternatives for Mobile Computers. *In Proceedings of the USENIX 1st Symposium on Operating Systems Design and Implementation (OSDI-94)*, Monterey, CA, USA, November 1994.
- [2] Linda Dailey Paulson, Will Hard Drives Finally Stop Shrinking? *IEEE Computer*, 38(5):14–16, May 2005.
- [3] A. Kawaguchi, S. Nishioka, and H. Motoda, A Flash Memory based File System. *In Proceedings of the USENIX Technical Conference*, 1995
- [4] M. L. Chiang, C. H. Paul, and R. C. Chang, Manage flash memory in personal communicate devices, *In Proceedings of IEEE International Symposium on Consumer Electronics*, 1997.
- [5] Li-Pin Chang, Tei-Wei Kuo, A Real-time Garbage Collection Mechanism for Flash Memory Storage System in Embedded Systems, *In Proceedings of the 8th International Conference on Real-Time Computing Systems and Applications (RTCSA 2002)*, Tokyo, Japan 2002.
- [6] K. Han-Joon, and L. Sang-goo, A New Flash Memory Management for Flash Storage System, *In Proceedings of the Computer Software and Applications Conference*, 1999.
- [7] Jen-Wei Hsieh, Li-Pin Chang, and Tei-Wei Kuo, Efficient On-line Identification of Hot Data for Flash-Memory Management, *In Proceedings of the ACM Symposium on Applied Computing (SAC)*, Santa Fe, New Mexico, USA, March 2005.
- [8] Sang-Won Lee and Bongki Moon, Design of Flash-Based DBMS: An In-Page Logging Approach, *In Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 55-66, Beijing, China, June, 2007.

SepRep: A Novel Reputation Evaluation Model in Peer-to-Peer Networks

Xiaowei Chen

Abstract

The computing model of networking system has changed from client/server to peer-to-peer. One of the most difficult and complex problems is the reputation computing and evaluation in pure distributed environment. We propose a robust and fast reputation model named SepRep in peer-to-peer networks. We give novel and brief definitions to clarify the reputation and trust for the first time, and discuss their different usage in the model. We take time factor into account and explain the reputation model into two parts: initial reputation and trust computing model, and reputation propagation model. Experimental results show that our model is robust against malicious peers, and it has fast convergence speed, even in a P2P environment with a large number of low trust malicious peers.

1 Introduction

Network development shows new trends: large scale distributing, global computing and global storage, etc. The requirement to the content access and transmission will be larger than to computing ability. Meanwhile, computing model of networking system has changed from C/S (Client/Server) to P2P (Peer-to-Peer) model. The large-scale application of P2P software (such as KaZaA, BitTorrent, eMule, Skype, etc.) provides strong evidences to the success of P2P computing model, especially in content distribution. Actually, the bandwidth consumption of P2P content distribution applications has already exceeded WWW (World Wide Web) and becomes the major part in Internet. Along with P2P's popularity, because of the anarchic nature of P2P systems, i.e. decentralized, autonomous and dynamic, etc., many related security problems are becoming more complex and different from C/S computing environment. One of the important issues is the influence of reputation and trust on P2P networks.

1.1 Background

The objective of reputation system in P2P network is to allow two sides of trade to judge the reliability of transaction by studying the peer's history behavior. According to Jøsang [1], the efficiency of trust mechanism should cover three factors: long time availability of attribution-entity object, acquirement and distribution of trust information, and decision-making by creditable information. Therefore, the main challenges focus on four parts:

- How to converge peers' reputation value with high speed and low overhead in distributed environment;
- How to ensure the accuracy of peer's reputation

value by evaluating or incorporating necessary information in P2P dynamic environment;

- How to detect or prevent the various attacks from malicious peers;
- How to ensure efficiency while the system scales up.

Currently, different reputation models or systems have been proposed to enhance the robustness, scalability and efficiency of P2P computing model, such as EigenTrust [2], P2Prep [3], Credence [4], NICE [5], PeerTrust [6], LIP [7], P-Grid [8] and PowerTrust [9], etc. They are typical and relatively successful reputation models. Amazon, eMule, eBay, Epinions and BitTorrent, etc. are typical and widely used practical P2P systems. In fact, the origin of reputation and trust in P2P comes from free-riding [10]. The recent research results show that, there are some bad behaviors in P2P system because of lacking efficient incentive and reputation mechanism [11]:

- Spreading virus, worm and Trojan horse [12].
- Fake file ratio is rising in P2P file-sharing network [13].
- Index poisoning in P2P file-sharing network [14].

1.2 Motivation

Though a number of reputation systems have been proposed and applied, some of them are based on centralized management, such as eBay, Amazon. It is not reliable and not applicable in pure distributed P2P environment. Some systems are based on distributed environment, but because of the complexity of P2P itself, there still exists challenges:

- How to enhance system's robustness against malicious peers;
- How to improve system convergence speed without increasing system overhead;
- How to make system more efficient in detecting misbehaviors.

Another important issue is the ambiguous concepts among reputation, trust and credibility. Some take them as the same meaning. Some take them different, give them differences, but there are still not obvious distinctions to these concepts in their models or systems.

We should take a new perspective to see what a peer really needs. In P2P content distribution, peer *i* wants to retrieve true and clean files with fastest speed, so it is apt to choose peers which can provide high quality of service (QoS). That is to say, in general, peer *i* always wants to receive the best service in P2P transactions. We can take this quality of service as reputation. In order to collect high QoS of peers, peer *i* updates reputations to other peers by transaction. But obviously, it is not efficient to update reputation only through direct transaction, so peer *i* can ask its neighbors about the

other side's reputation to speed up the updating. Of course, the more creditable the neighbors are, the more accurate the reputations peer i can get, and the reputations will also approach the real value faster. Here we can see the trust is to help peer to get more correct reputation with faster convergence speed. Reputation and trust complement each other. Therefore, we need to resolve is that how to make them complement better.

In this paper, we separate, clarify and redefine the concepts of reputation and trust with a new vision, eliminate the possibility of different meanings. We propose a hybrid reputation and trust overlay network (HRTON) to evaluate each peer's reputation from both objective and subjective aspects, then present our reputation evaluation and propagation model named SepRep, which is based on pure distributed environment. The experimental result validates that SepRep reputation model can converge quickly and enhance system robustness against malicious peers efficiently. It can distinguish malicious peers with low overhead in P2P networks.

The remaining parts of this paper are organized as follows: Section 2 reviews existing work on P2P reputation systems. In Section 3, we introduce the basic SepRep reputation model and give related definition. The approach of getting the initial reputation and reputation propagation model will be explained in detail. We evaluate the performance of our model and analyze the experimental results in section 4. Finally, conclusion and suggestions for future work will be summarized.

2. Related Works

Many literatures try to exactly define the concepts of reputation and trust. Due to the universality of the concepts, the understandings of them appear diversity. According to the ITU-T X.509, Section 3.3.54, trust is defined as follows: "Generally an entity can be said to 'trust' a second entity when the first entity makes the assumption that the second entity will behave exactly as the first entity expects." That means, trust is an indicator of credibility to content, and it is comparable. Another very similar concept is reputation. According to a formal definition of reputation given by Wilson [15], together with P2P environment, it is "a characteristic or attribute ascribed to one peer (or peers) A by another person (or peers) B". On the other hand, the reputation is also considered as a service provider which can be formed by means of a collection of ratings by different users, each such rating is intuitively equivalent to user satisfaction. Besides, Jøpsang distinguishes the trust and reputation: trust is divided into reliability trust and decision trust, and reputation is viewed as a collective measure of trustworthiness based on the referrals or ratings from members in a community [1].

Though reputation and trust have various definitions, they are interrelated tightly, have some common features, so some researchers take the two concepts as the same meaning. But the biggest difference between reputation and trust is that reputation is an objective

concept, and that trust is a subjective concept. We can use one sentence to describe them: I trust you because you have good reputation; I trust you despite your bad reputation [1].

There are some new reputation systems in P2P in recent years. They provide different approaches to evaluate reputation.

Credence is a subjective, independent and local reputation mechanism based on Gnutella. It defines polling mechanism, which let users vote for whether the sharing file matches the file description or not. Credence exchanges reputation table among the selected high reputation value users, extends reputation relationship by reputation transitivity, and chooses the path with highest reputation value as the peer's reputation value. Credence uses file as the basis of building reputation relationship. It can avoid dynamic feature of users' behaviors and can judge the essential attribution of file. But peer's reputation value will be affected in users' voting because of users' subjectivity, especially in collusion attack. And it needs to solve the problem about how to prompt users' spontaneity.

TrustGuard is a secure reputation mechanism framework based on PeerTrust. The major goal of TrustGuard focuses on the vulnerabilities of a reputation system itself. The authors identify three types of threats, that is, strategic oscillations, fake transaction and dishonest feedback, and provide corresponding countermeasures. In this framework, each peer has a transaction management unit, a reputation evaluation engine and a feedback data storage unit. The three components' computing uses strategic oscillation guard and dishonest feedback guard to ensure the correctness. TrustGuard uses modular design, it does not need to worry about the other parts of system when adding new guard module or modifying current module.

LIP is an objective, global reputation mechanism. LIP discovered and proved that "users are apt to remain real files in a long time, and delete polluted files in a short time". It gives statistics automatically about file's remaining time in each user's computer, and then computes the number of holders to each file and the file's average remaining time in user's computer. The objective statistics feature of LIP can make it get more reliable information, and it can collect complete information without adopting incentive mechanism.

PowerTrust is a global, robust and scalable reputation system based on power-law. It uses trust overlay network (TON) model to analyze the power-law distribution of peer feedbacks. The system offers very fast global reputation aggregation, ranking and updating, together with robustness and wide applicability. PowerTrust does not solve collusion problem well and it has not supported unstructured P2P system currently.

3. SepRep Model Methodology

SepRep model offers clear definitions to reputation and trust to evaluate peer's related performance and behavior. We use hybrid reputation and trust overlay

network (HRTON) to connect local reputation and local trust together. In this section, we introduce the system concept and discuss new features in SepRep.

3.1 The SepRep Model Concept

Although we can get a difference between reputation and trust, we still can not distinguish them in P2P networks. Actually, the difference of objective and subjective to reputation and trust inspired us to offer them a novel and reasonable representation. Our proposed mechanism will make use of positive and negative information to rate a peer, both from own and from other's experience.

Definition 1 (Reputation). *Reputation is an objective concept. It is a quantified QoS (Quality of Service) description hold by a peer to another peer in P2P networks. The QoS description is formed by all the aspects of service quality which a peer can provide. It can be divided into direct reputation, indirect reputation, local reputation and global reputation.*

Here reputation is an objective concept, and QoS can be evaluated with a peer's bandwidth (both the download and upload), the number of sharing contents owned by a peer, content validity (i.e., the relativity between file name and file content, banding with virus or advertisement software or not, etc.), the online duration time of a peer, etc. Reputation can be calculated by centralized server, i.e., eBay, Amazon, Epinions, etc. It also can be calculated by each distributed peer. In fact, reputation in P2P CDN (Content Distribution Network) should not be calculated by centralized server, but by distributed peer.

There are *direct reputation (DR)*, *indirect reputation (IR)*, *local reputation (R)* and *global reputation (GR)* in our system. When peer i has a transaction with peer j , the direct reputation $DR_{i \rightarrow j}$ represents the direct opinion of peer i on peer j 's behavior in the transaction. The indirect reputation $IR_{i \rightarrow j}$ represents the opinions collected by peer i from other peers on peer j 's behavior. The local reputation $R_{i \rightarrow j}$ represents the opinion of peer i on peer j 's behavior in the P2P system. The GR of a peer is the average value which is calculated by the sum of local reputation of the peer. In pure distributed environment, actually there is no global reputation stored in each peer, but we can calculate it.

Definition 2 (Trust). *Trust is a subjective concept. It is quantified credibility hold by a peer to another peer in P2P networks. The credibility represents the opinion of a peer on how honest another peer is in the distributed computing of the reputations.*

In our approach, all peers are labeled from 1 to n . Each peer i maintains two rating vector, namely, the reputation rating $R[R_{i \rightarrow 1}, R_{i \rightarrow 2}, \dots, R_{i \rightarrow n}]$ and the trust rating $T[T_{i \rightarrow 1}, T_{i \rightarrow 2}, \dots, T_{i \rightarrow n}]$. $R_{i \rightarrow j}$ or $T_{i \rightarrow j}$ means the reputation or trust of peer i on peer j . We can clearly get a hybrid reputation and trust overlay network (HRTON) in P2P system. It is a virtual relation network, shown in figure 1.

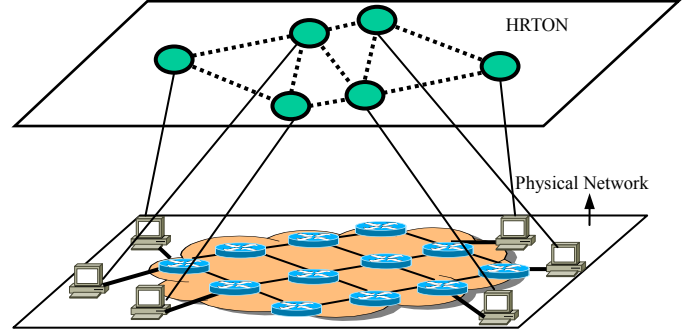


Fig. 1: Hybrid Reputation and Trust Overlay Network

All the calculation and propagation are based on the network. Then we get two matrices, reputation matrix M_R and trust matrix T_R , just as follows:

$$M_R = \begin{bmatrix} R_{1 \rightarrow 1} & R_{1 \rightarrow 2} & \dots & R_{1 \rightarrow n} \\ R_{2 \rightarrow 1} & R_{2 \rightarrow 2} & \dots & R_{2 \rightarrow n} \\ \dots & \dots & \dots & \dots \\ R_{n \rightarrow 1} & R_{n \rightarrow 2} & \dots & R_{n \rightarrow n} \end{bmatrix}$$

$$T_R = \begin{bmatrix} T_{1 \rightarrow 1} & T_{1 \rightarrow 2} & \dots & T_{1 \rightarrow n} \\ T_{2 \rightarrow 1} & T_{2 \rightarrow 2} & \dots & T_{2 \rightarrow n} \\ \dots & \dots & \dots & \dots \\ T_{n \rightarrow 1} & T_{n \rightarrow 2} & \dots & T_{n \rightarrow n} \end{bmatrix}$$

Typically, we assume that $0 \leq R_{i \rightarrow j} \leq 1$: the higher the value of this number, the better peer i views peer j , while $R_{i \rightarrow i}$ has no meaning in SepRep, because any peer i can not judge itself objectively. We also assume that $0 \leq T_{i \rightarrow j} \leq 1$: $T_{i \rightarrow j} = 0$ denotes that peer i does not trust any information from peer j at all, while $T_{i \rightarrow j} = 1$ denotes that peer i trusts everything from peer j . Obviously, $T_{i \rightarrow i} = 1$ for any peer i .

3.2 Initial Reputation and Trust Model

In HRTON, all peers are not familiar with each other at the beginning. Their reputation and trust will be assumed the same value. Since the range of reputation and trust is between 0 and 1, we can assign an intermediate value to each peer.

True value of reputation and trust will be assigned several rank values instead of gradient values in each transaction, i.e., reputation $\in \{\text{low, middle, high}\}$, trust $\in \{\text{distrust, uncertain, trust}\}$. The rank can be more than three grades. It depends on the practical requirement. For simplicity, we don't take the middle or the uncertain rank into account at the initial phase (they will be considered in our propagation model), so we can classify all peers as four kinds in HRTON:

- HRHT: Peer has high reputation value and high trust value. It means we prefer transacting with this peer and prefer trusting information or recommendation from this peer.

- HRLT: Peer has high reputation value and low trust value. It means we prefer transacting with this peer but not prefer trusting information or recommendation from this peer. It can be viewed as a malicious peer. This kind of malicious peer behaves well during transactions in order to get high normalized reputation values as seen by other peers, but it always reports false local reputation values of other peers (i.e. too high or too low). By combining good behaviors with reporting false local reputation values, a malicious peer can thus cause strategic oscillation attacks in P2P networks.
- LRHT: Peer has low reputation value and high trust value. It means we prefer not transacting with this peer but prefer trusting information or recommendation from this peer.
- LRLT: Peer has low reputation value and low trust value. It means we prefer not transacting with this peer and not trusting information or recommendation from this peer. It can be viewed as another kind of malicious peer. This kind of peer always behaves not well for the limitation of itself condition and reports false local reputation values of other peers. It can cooperate with other peers to process collusion attack or may perform “whitewash” operation to shed previous bad reputation.

Then we will explain the computing model in detail. First, we consider a simple situation. In one transaction, there exist three kinds of peers: transaction receiver peer i, transaction sender j, transaction agency peer k. Figure 2 is the sketch map.

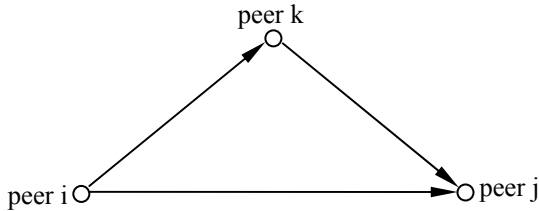


Fig. 2: Transaction Sketch Map

Peer i will request transaction with peer j, it will give evaluation to j according to j’s QoS. If necessary, it will consider the peer k’s suggestion to judge whether it should transact with j or not.

Initial reputation computing will combine the direct transaction evaluation and indirect evaluations from agency peers. Direct reputation is calculated as follows:

$$DR_{i \rightarrow j}^{<n>} = \alpha(t)DR_{i \rightarrow j}^{<n-1>} + (1 - \alpha(t))R_{Evaluation}^{<n>} \quad (1)$$

$DR_{i \rightarrow j}^{<n>}$ means the n-th direct reputation of i on j.

The above formula means the current direct reputation will be decided by the last direct reputation, time factor $\alpha(t)$ and the current evaluation $R_{Evaluation}$. Obviously, if the transaction scale is the same, then the reputation of a peer in a transaction a year ago should be lower than in a transaction a week ago, so time is the necessary factor.

Next we will compute the indirect reputation. Peer i

will ask peer k to about peer j’s reputation. Peer k’s trust will be counted in it. It is calculated as follows:

$$IR_{i \rightarrow j} = T_{i \rightarrow k} \cdot R_{k \rightarrow j} \quad (2)$$

Apparently, it is not sufficient to ask only peer j, so we will ask multiple peers. Here we define a set $K^{<n>}$ to denote peer i’s random neighbors. Instead of simple addition, we adopt a normalized integration as follows:

$$IR_{i \rightarrow j} = \sum_{k \in K^{<n>}} R_{k \rightarrow j} \cdot t_{i \rightarrow k} \quad (3)$$

$$t_{i \rightarrow k} = T_{i \rightarrow k} / \sum_{k \in K^{<n>}} T_{i \rightarrow k} \quad (4)$$

After we get $DR_{i \rightarrow j}$ and $IR_{i \rightarrow j}$, we can calculate local reputation $R_{i \rightarrow j}$ as follows:

$$R_{i \rightarrow j} = \beta DR_{i \rightarrow j} + (1 - \beta) \cdot IR_{i \rightarrow j} \quad (5)$$

β is a weight balance parameter. It denotes the proportion of direct reputation and indirect reputation. It lies on a peer prefer trusting itself transaction records or other peers’ recommendation reputation values.

Global reputation GR_i can be easily calculated as follows. It is an average of reputation value of i given by peer k, and peer k is a peer who has transaction with peer i.

$$GR_i = \sum R_{k \rightarrow i} / \sum_{k=1}^n k \quad (k \neq i) \quad (6)$$

Then we will calculate the trust value. According to the trust definition given in this model, trust is the opinion of peer on how honest another peer is in the distributed computing of the reputations, peer i only can give j’s reputation evaluation, but not the trust. We can use the similar way to get the peer i to peer j’s trust. We can calculate peer k’s trust value in this transaction. During the transaction between peer i and j, we can get the i to k’s indirect trust by asking j about k’s trust. It is calculated as follows:

$$IT_{i \rightarrow k} = T_{j \rightarrow k} \cdot T_{i \rightarrow j} \quad (7)$$

Then the local trust $T_{i \rightarrow k}$ is calculated as follows:

$$T_{i \rightarrow k}^{<n>} = \gamma(t)T_{i \rightarrow k}^{<n-1>} + (1 - \gamma(t)) \cdot IT_{i \rightarrow k}^{<n>} \quad (8)$$

$\gamma(t)$ has the same purpose as $\alpha(t)$. In order to fasten the convergence speed, we add feedback function to make the trust value reach the real value faster. So the trust value is updated as follows:

$$T'_{i \rightarrow k} = f(T_{i \rightarrow k}, |R_{i \rightarrow j} - R_{k \rightarrow j}|) \quad (9)$$

The feedback function is to update $T_{i \rightarrow k}$ according to the difference between $R_{i \rightarrow j}$ and $R_{k \rightarrow j}$. The bigger the difference value is, the smaller the $T_{i \rightarrow k}$ will be.

3.3 Reputation Propagation Model

We then discuss our reputation propagation model. Based on the above explanation, the second hand reputation is defined as $R_{i \rightarrow j}^{<2>} = \sum_k R_{k \rightarrow j} \cdot t_{i \rightarrow k}$. i.e., $R^{<2>} = T \cdot R$. Similarly, the h-th reputation is defined as follows:

$$R^{<h>} = T \cdot R^{<h-1>} \quad (h > 2) \quad (10)$$

This approach guarantees that the learned reputation $R_{i \rightarrow k}^{<h>}$ of peer i on any peer k (via at most h-hop queries) is bounded by $\max R_{i \rightarrow k}$.

This reputation model has the ‘‘loss goes shares’’ effect. Let peer j be a malicious peer, it will offer high quality services with 10% probability. If peer i request direct transaction with j, then the interactive times will be 10 at least to get j’s correct reputation value, and i will get 9 low quality services (or suffer losses) in the 10 transactions. But taking our approach, i only need to use information provided by 10 trust buddies who have transaction history with j, it can get the same correct result. The difference is the losses goes shares by other 9 peers. The meaning is that it is easier to find 10 peers that each of them has one transaction with j than to find 1 peer which has 10 transactions with j.

4. Experiment and Analysis

4.1 Simulation

Considering the social network in real world, people tend to distrust more than trust at first. So we can assign a value lower than 0.5 to each peer as its reputation and trust value, i.e., around 0.4. In HRTON, the initial real reputation and trust value configurations list in table 1, and experiment configurations list in table 2.

Table 1: HRTON Initial Real Value Table

Peer Type	Real Value
High Reputation (HR)	[0.75..0.85]
Low Reputation (LR)	[0.15..0.25]
High Trust (HT)	[0.75..0.85]
Low Trust (LT)	[0.15..0.25]
Random	[0..1]

Table 2: Experiment Configuration Table
(‘‘Rep’’ indicates ‘‘Reputation’’)

Type	Rep Type	Rep Ratio	Trust Type	Trust Ratio
1	HR	30%	HT	80%
	LR	70%	LT	20%
2	HR	30%	HT	50%
	LR	70%	LT	50%
3	HR	30%	HT	20%
	LR	70%	LT	80%
4	Random	100%	Random	100%

As shown in table 1, we give peer’s initial real reputation and trust value. High value is around 0.8, low value is around 0.2, while the random value is designed for experiment type 4. In table 2, we design four types of experiment. According to Adar [10], large proportion of the user population, upwards of 70%, enjoy the benefits of the system without contributing to its content. So in type 1-3, the high reputation ratio keeps 30%, and the low reputation ratio keeps 70%. In type 4, we assign random reputation and trust values to peers in order to evaluate the model’s convergence speed in the general situation.

Peer with high trust value can be considered as good peer who always reports correct reputation value to other side, and peer with low trust value can be considered as malicious peer or bad peer. Here malicious behavior means the peer report a random value to the other peer in a transaction, whatever the other peer has good or bad reputation.

Table 3: Experiment Parameters

Parameters	Value
Peer number	500
Iteration time	5000
$\alpha(t)$	0.1
β	0.6
$\gamma(t)$	0.1
Neighbor number of a peer	20

In our simulation, there are 500 peers which process 5000 iterations. Here iteration indicates that each peer will execute a transaction with another random peer in one iteration. That means a peer will execute 5000 transactions to other peers. In each transaction, peer i will ask 20 neighbor-peers to get the indirect reputation. In order to decrease the effect of malicious peers report incorrect reputation values, we directly neglect the report reputation values from neighbor peers which trust values are lower than 0.25. About the feedback function to trust, we will give a feedback value according to the difference between $R_{i \rightarrow j}$ and $R_{k \rightarrow j}$. The bigger the difference value is, the smaller the feedback value will be. Then we take the feedback value as a plus trust evaluation to update the current trust value. It is calculated as follows:

$$T'_{i \rightarrow k} = \frac{\sum_{transaction} T_{i \rightarrow k} + feedback}{\sum_{transaction} + 1} \quad (11)$$

4.2 Analysis

Here we use error sum of squares (SSE) to indicate the peer’s reputation deviation to the real value in each iteration, shown in formula 12, and the 5000th iteration deviation value is shown in table 3. Rep_{real} indicates peer i’s real reputation value, and $Rep_{compute}$ indicates peer i’s reputation value which is computed by SepRep

model (i is the identifier of a peer, $i \in [1..n]$).

$$Error_{iteration} = \sqrt{\sum_{i=1}^n (Rep_{real} - Rep_{compute})^2} \quad (12)$$

Table 3: Reputation SSE in 5000th Iteration

Experiment Type	Error Sum of Squares
1	0.0813335
2	0.110152
3	0.23338
4	0.167947

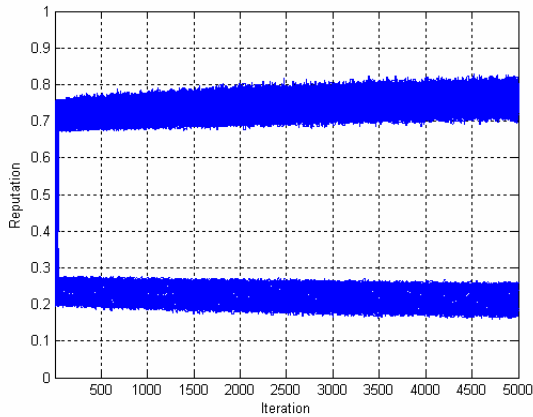


Fig. 3: Experiment Type 1

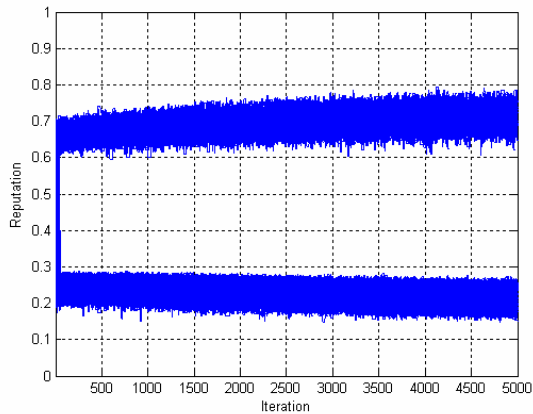


Fig. 4: Experiment Type 2

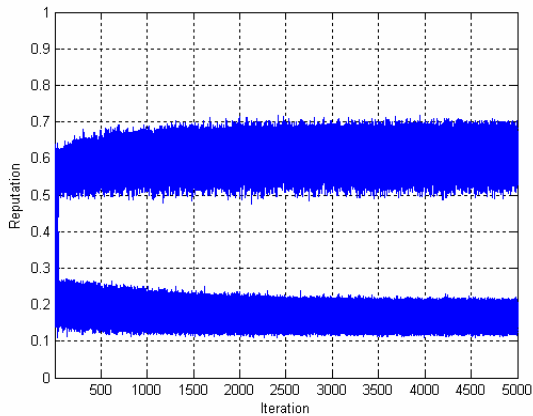


Fig. 5: Experiment Type 3

Figure 3-5 show the peer reputation values change differently in experiment of type 1-3, and convergence speed is shown in figure 6. In the figure 3-5, we can see that our model can distinguish high and low reputation peers very quickly, and then converge to the real value. Figure 3 shows that a P2P network has 80% credible peers, the high reputation peers change from initial value to the final steady status (very close to the real value 0.75 to 0.85). Figure 4 shows that a P2P network has 50% credible peers, and Figure 5 shows that a P2P network has 20% credible peers. Along with the increasing of low trust peers, reputation value will offset more and more from the real value, but the offset value keeps a relative reasonable range, and the convergence speed keeps the same. That means the malicious peers' increasing will affect the accuracy of reputation and trust value. The more malicious peers exist, the more inaccuracy reputation and trust values get. In figure 5, the error offset to real value is over 0.2 at the 5000th iteration (shown in table 3). Though the reputation value is not correct to the real value, but the offset is acceptable in the worst situation, and we still can distinguish good and bad peers very clearly.

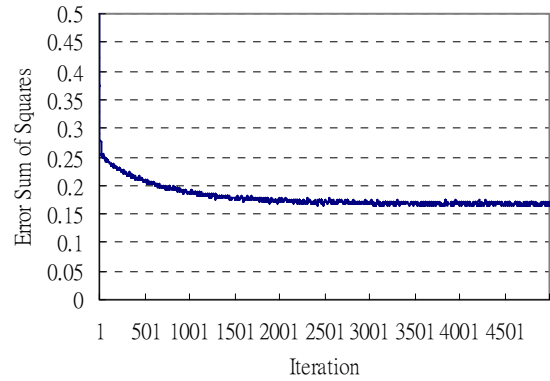


Fig. 6: Reputation Convergence Speed in Experiment Type 4

According to the figure 6, we can clearly see that reputation will converge to a steady status after the 2000th iteration. But in the steady status, there still exists error (0.168) to the real value, and this error value is lower than the experiment type 4 which indicates the worst situation in P2P networks. Because in the experiment of type 4, all the peers' reputation and trust values are random, it represents a general situation, so this error is acceptable and reasonable.

5. Conclusion and Future Work

5.1 Conclusion

We propose a robust and fast reputation model SepRep in P2P networks. We redefine the reputation and trust at first, and discuss their different usages in the model. Then we explained the reputation model into two parts: initial reputation and trust computing model,

and reputation propagation model. In this model, we take time factor into account, use direct reputation and indirect reputation to get the local reputation. Experimental results show that our model has fast convergence speed and robustness to malicious peers.

5.2 Future Work

Concerning about the limitation of reputation that peers in P2P network often focus on some certain fields, such as music, movie, software, etc., which means peer's interest will be different. We plan to propose reputation interest group (RIG) to achieve more accurate reputation value according to peers' different content interest. This also may increase the distribution efficiency in P2P environment.

For the content distribution environment, we will classify content as four different parts: video, audio, software and other materials (such as e-book, teaching materials, etc.). Each group can overlay the others. As shown in figure 7, four circles indicate four kinds of content, red area indicates type 1, purple area indicates type 2, green area indicates type 3, and blue area indicates type 4.

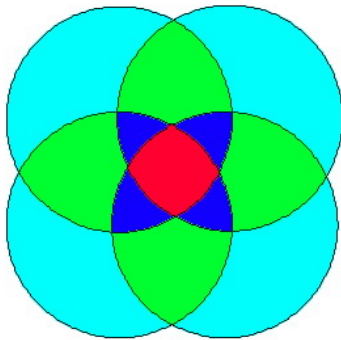


Fig. 7. Reputation Interest Group

Peers will select their interest to perform content distribution. In order to decrease the system overhead and compatible with SepRep algorithm, each peer's reputation will be calculated as one value.

We also intend to solve whitewashers and Byzantine problems in SepRep in future. Solving of them will further increase the robustness of our model.

References

- [1] A. Jøsang, R. Ismail and C. Boyd, "A Survey of Trust and Reputation Systems for Online Service Provision," *Decision Support Systems*, 2005.
- [2] S. Kamvar, M. Schlosser, "The EigenTrust Algorithm for Reputation Management in P2P Networks," *WWW*, Budapest, Hungary. 2003.
- [3] F. Cornelli, E. Damiani and S. De Capitani, "Choosing Reputable Servents in a P2P Network," *In Proceedings of the 11th World Wide Web Conference*, Hawaii, USA, 2002.
- [4] K. Walsh and E. G. Sirer, "Fighting peer-to-peer SPAM and decoys with object reputation," *Applications, Technologies, Architectures, and Protocols for Computer Communication*, pp. 138-143, 2005.
- [5] S. Lee, R. Sherwood, "Cooperative peer groups in NICE," *IEEE Infocom*, San Francisco, USA. 2003.
- [6] L. Xiong, L. Liu, "A Reputation-Based Trust Model for Peer-to-Peer eCommerce Communities," *IEEE International Conference on E-Commerce (CEC)*, 2003.
- [7] Q. Feng and Y. Dai, "LIP: A Lifetime and Popularity Based Ranking Approach to Filter out Fake Files in P2P File Sharing Systems," *Peking University*, 2006.
- [8] Karl Aberer, Zoran Despotovic, "Managing trust in a peer-2-peer information system," *10th Intl Conference on Information and Knowledge Management (CIKM)*, Atlanta, 2001.
- [9] Zhou R., Hwang, K., "PowerTrust: A Robust and Scalable Reputation System for Trusted Peer-to-Peer Computing", *IEEE Transactions on Parallel and Distributed Systems*, Vol. 18, No. 4, pp. 460-473, April 2007.
- [10] E. Adar, Free Riding on Gnutella, 2000. <http://www.hpl.hp.com/research/idl/papers/gnutella/gnutella.pdf>
- [11] Qin-yuan Feng, Ya-fei Dai, "P2P network trust mechanism review," *Communication of CCF*, pp.31-40, Mar. 2007.
- [12] L. Zhou, L. Zhang, F. McSherry, N. Immorlica, M. Costa and S. Chien, "A First Look at Peer-to-Peer Worms: Threats and Defenses," *Proceedings of 4th International Workshop on Peer-to-Peer Systems (IPTPS)*, 2005.
- [13] N. Christin, A. S. Weigend and J. Chuang, "Content availability, pollution and poisoning in file sharing peer-to-peer networks," *Proceedings of the 6th ACM conference on Electronic commerce*, pp. 68-77, 2005.
- [14] J. Liang, N. Naoumov and K.W.Ross, "The Index Poisoning Attack in P2P File-Sharing Systems," *In Proceedings of IEEE Infocom*, 2006.
- [15] Wilson, R., "Reputation in games and markets," In A. Roth (Ed.), *Game-theoretic models of bargaining*, pp 65-84, *New York: Cambridge University Press*, 1985

Performance Enhancement and Simulation Implementation of IEEE 802.11 Infrastructure Network

Yan Yong

Abstract

To meet the increasingly demand of ubiquitous internet service, the IEEE 802.11 wireless local area network (WLAN) standard has been widely deployed under its infrastructure mode, which mandates that all packets transmissions must be forwarded at Access Points (AP). Due to this nature of 802.11 infrastructure mode, a problem will occur inevitably when there is intensive intra-cell connectionless traffic, such as UDP, present in the network, since the packets of the connectionless traffic will eventually overflow the AP buffer as well as occupy the air channel, which will freeze the connection-oriented service. This problem greatly weakens the IEEE 802.11 infrastructure mode and, as a consequence, brings serious unfairness problem to other connection-oriented services such as the Transport Control Protocol.

In this paper we point out this inherent problem, and propose several possible solutions against this problem to enhance the original 802.11 protocol. Furthermore, this paper studies the simulation implementation of the IEEE 802.11 infrastructure mode on the widely recognized platform Network Simulator 2 (ns-2), in order to realize and verify the functionality of our proposed solutions.

1. Introduction

IEEE 802.11 protocol [4] supports two kinds of operation modes, as shown in Figure 1: (1) In the infrastructure mode, an Access Point (AP) acts as a central node. A station has to send its message to AP first, while the AP is responsible for forwarding all messages to the destinations, which can be either inside or outside of this wireless cell; (2) In ad hoc mode, there is no AP in the network and the wireless stations are able to communicate with each other directly. The infrastructure mode is the dominant mode widely adopted by most of the current wireless LANs, where AP performs the role of a portal (router) to Internet, as well as a central node of the local wireless cell.

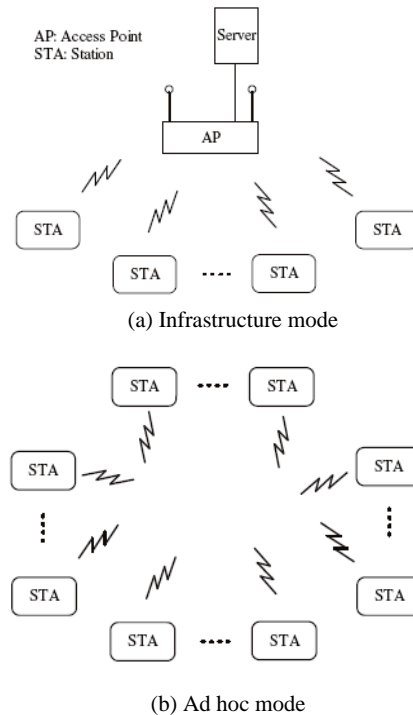


Figure 1: Infrastructure mode versus ad hoc mode

The basic access method in the 802.11 MAC protocol is DCF (Distributed Coordination Function) based on the Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) [4]. DCF employs a distributed CSMA/CA algorithm and an optional virtual carrier sense using RTS and CTS control frames. When using the DCF, before initiating a transmission, a station senses the channel to determine whether another station is transmitting. If the medium is found to be idle for an interval that exceeds the Distributed InterFrame Space (DIFS), the station proceeds with its transmission. However if the medium is busy, the transmission is deferred until the ongoing transmission terminates. A random interval, henceforth referred to as the backoff interval, is then selected; and used to initialize a backoff timer. The backoff timer is decreased as long as the channel is sensed idle, stopped when a transmission is detected on the channel, and reactivated when the channel is sensed idle again for more than a DIFS. The station transmits when the backoff timer reaches zero. When more than one node are counting down their backoff timers simultaneously, there is a probability that some of them have their timers reach zero at the same time slot, and start transmitting at the

beginning of next time slot exactly at the same time, which results in a collision.

2. The problem with 802.11 infrastructure mode

The performance of 802.11 DCF has been studied in the literature through analytical models, simulations, and experiments [8-17]. It is well known that DCF throughput degrades gracefully under increasing multiple access contention. However, previous study always assumes that AP is just a bridge between the local network and the outside network. As far as we know, there is no systematic study about the system performance when intra-cell UDP traffic exists.

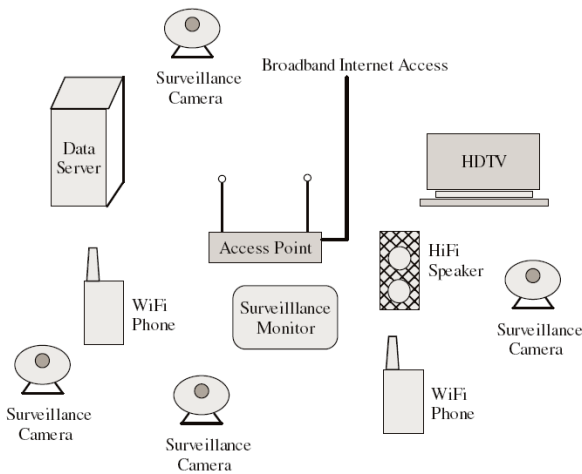


Figure 2: a typical scenario of digital home based on a wireless LAN. A typical scenario with intensive intra-cell UDP traffic is shown in Figure 2, which simulates an environment of a digital home employing IEEE 802.11 wireless LAN. In this scenario, lots of real time applications such as wireless surveillance system, voice over WiFi, HiFi over WiFi, as well as the Internet access service, are all supported by a wireless LAN with an AP.

Given one AP and n stations and at saturation condition, there are always $1 + n$ senders, and hence the AP has only $1 / (1 + n)$ of chance to capture the medium.

Let's demonstrate the above theory through the simplest case. As shown in Figure 3, in the infrastructure mode, if there are two stations keep on sending packets to each other through the AP, the chance that AP seizes the channel is roughly $1/3$, while the two stations get the remaining $2/3$ bandwidth. Hence, the AP's buffer will be full after a short time. At the moment when the buffer is full, the AP starts to drop packets.

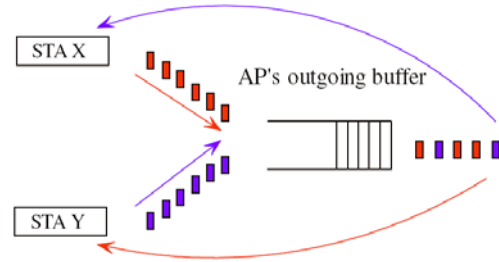


Figure 3: AP buffer is overflowed by connectionless traffic

As a result of this problem, other connection-oriented services, if existing, will freeze, as the AP will not be able to deliver any ACK to them after its buffer is full. In other words, due to the connectionless traffic senders' greedy or careless behavior, the connection-oriented service will sacrifice, which causes a serious fairness issue. Unfortunately, this problem is inevitable with the current 802.11 infrastructure mode, when intensive intra-cell UDP traffic appears.

3. Possible Solutions

There are several possible ways to solve the problem, we briefly list some of them:

(1) To implement Direct Link Protocol proposed in 802.11e standard [6] in all the wireless stations and AP. This solution has the highest efficiency, but it does not work for a wireless LAN in which some legacy wireless stations exist.

(2) To make it an atomic operation for the AP to relay an intra-cell packet: once the AP has received a packet destined for another wireless station in the same wireless LAN, after sending the ACK to the source as usual, the AP waits for an SIFS and then forwards the packet to the destination.

(3) To design rate control scheme at the MAC layer. If the AP's buffer is almost full, the AP notifies the wireless stations through the ACK frames. The wireless stations receiving such notifications shall slow down their sending rate. This solution needs to slightly modify the format the ACK frame, and it has the same drawback as solution (1).

(4) To design rate control scheme at the application layer. We limit our discussion to UDP applications only. Without the supporting from MAC layer, the application layer has to detect buffer overflow at the AP by itself and then react accordingly.

4. Simulating wireless 802.11 networks

4.1 NS2 Background

Network Simulator 2 (ns2), is a discrete event driven simulator for networking protocols. It evolved from REAL variant (1989) and DARPA (LBL, Xerox PARC,

UCB, and USC/ISI) (1995). Ns2 is one of the most popular network simulators and is open source. Ns2 is an object oriented, written in C++, with an OTcl interpreter as a frontend. The simulator supports a class hierarchy in C++, and a similar class hierarchy within the OTcl interpreter. The two hierarchies are closely related to each other; from the user's perspective, there is a one-to-one correspondence between a class in the OTcl hierarchy and one in the C++ hierarchy. TclCL (Tcl with classes) is a Tcl/C++ interface used by Mash, vic, vat, rtp_play, ns, and nam. It provides a layer of C++ glue over OTcl. Figure 4 shows the relationship of OTcl and C++ hierarchy, and the TclCL interface.

4.2 Linkage of C++ and OTcl

C++ is fast to run but slower to change, making it suitable for detailed protocol implementation. OTcl runs much slower but can be changed very quickly (and interactively), making it ideal for simulation configuration.

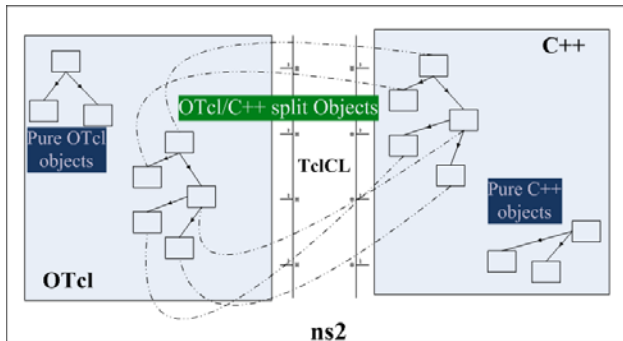


Figure 4: Relationship of OTcl C++ and TclCL

Basically, OTcl is used when:

- for configuration, setup, and “one-time” stuff
- the new code can do what it wants by manipulating existing C++ objects

and C++ should be used when:

- the new code is doing *anything* that requires processing each packet of a flow
- the new code has to change the behavior of an existing C++ class in ways that weren't anticipated

In general, if the new code needs to invoke Tcl many times per second, probably it's necessary to move that code to C++ hierarchy.

Linkage from C++ to OTcl:

OTcl creates objects in the simulator

Objects are *shared* between OTcl and C++ by default

- Access the OTcl interpreter via class Tcl
- `Tcl &tcl = Tcl::instance();`
`tcl.eval(...); tcl.evalc(“”); tcl.evalf(“”,...);`
`tcl.result(...); res = tcl.result();`
- ns2 also keeps hashtable of every TclObject
`tcl.enter()/lookup()/remove()`

Variables are, by default, *unshared*

- Pure C++ or OTcl variables
- Bindings can be established by the compiled constructor
`bind(); bind_delay();`
{ `bind(“rtt_”, &t_rtt_); t_rtt_ = 10; }`

Linkage from OTcl to C++

- For all TclObject, ns creates `cmd{}` instance procedure to access compiled methods
- Consider `$o distance <agentaddr>`
 - Invokes `distance{}` instance procedure of `$o`
- If this doesn't exist, call parent TclObject `unknown{}` method
- ...which calls `$o cmd distance <agentaddr>`
- ...which calls C++ `<T>::command()` method

Four major types of ns2 components

- Application
 - Communication instigator
- Agent
 - Packet generator/consumer
- Node
 - Addressable entity
- Link
 - Set of queues

4.3 Wireless Simulation

Ns2 supports 802.11 link as well as mobile ad hoc network. The basic mobile node architecture is based on the CMU Monarch extension to ns, as shown in Figure 5. However, this extension does not focus on the wireless network, but on the mobility support with mobile IP and mobile node.

The Mac IEEE802.11 in NS-2 implements:

- DCF implemented by Monarch project
- Basic Access
- RTS/CTS/DATA/ACK
- Backoff procedure
- Carrier-sense mechanism
- Both WLAN & Ad Hoc
- Physical layer not well implemented

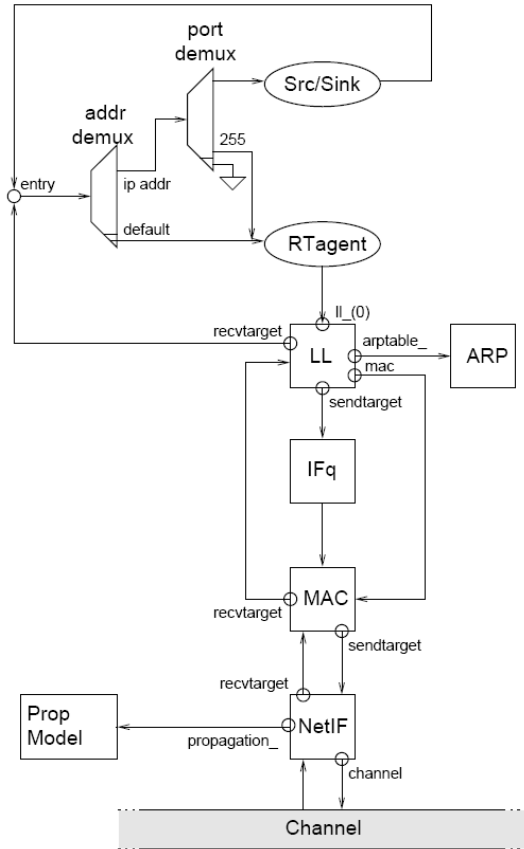


Figure 5: Schematic of a mobile node under the CMU Monarch wireless extensions to ns [2]

4.4 Mobile node Component Detail [2]

Here We list the components detail of mobile node in Figure 5.

4.4.1 Classifier/Addr

It's a bit of a hack, but if the `shift_` field is set to 0, the classifier will demux based on the destination port number, and if the `shift_` field is set to 8, the classifier will demux based on the destination IP address. Other values for `shift_` will cause an abort.

4.4.2 class LL

The class LL is based on code by the Daedalus Research Group at UCB [18].

4.4.3 class Mac

All objects implementing a MAC protocol are derived from class Mac. The Mac object will normally pass to its `recvtarget` only packets matching the id set for the object. Agents that subclass themselves from class Tap defined in `cmu/mac.h` can also register themselves with the Mac object using `installTap()`. If the particular MAC protocol in use permits it, the tap will be promiscuously given all packets received by the Mac layer, before address filtering is done.

4.4.4 class PriQueue

Of Tcl type Queue/DropTail/PriQueue, class PriQueue inserts routing protocol packets at the head of the queue, and all other packets at the back. It also supports running a filter over all the packets in the queue and removing those with a specified destination address.

5. 4.5 class NetIf

class NetIf is the superclass from which all network interfaces must be subclassed. Currently we provide only NetIf/SharedMedia interfaces.

4.4.5.1 NetIf/SharedMedia

Implements a generic radio interface characterized by the property that it behaves as a shared media between nodes. What one node transmits, the other nodes can receive (subject to the propagation model, of course).

4.5 IEEE 802.11 implementation in ns-2 [19]

4.5.1 WirelessChannel

The WirelessChannel simulates the physical media (air) for wireless communication. The class WirelessChannel inherits from class Channel. The implementation is in ns2/mac/channel{.cc,.h} .

The WirelessChannel keeps a list of all nodes on this channel. The list is sorted based on the X-dimension values of nodes before it can be used. A state delay_ is kept by the channel (inherited from Channel).

All functionalities about the contention, resume, transmission in the implementation of Channel in previous version of ns2 have been removed to /mac. So even WirelessChannel inherits from the Channel, it works just like a physical media: receiving packets from nodes and assigning them to all possible destination (nodes in range of the sender).

4.5.2 MobileNode

The class MobileNode simulates the mobile or wireless nodes in wireless ad hoc networks or sensor networks on wireless channel. It inherits from the class Node. The main difference is that there is no links on any MobileNode, the receiving and sending packets are based on the WirelessChannel not the links. It also simulates the mobility of nodes, based on the speed and destination locations.

The implementation is at /common/mobilenode{.cc,.h} . Mostly, class MobileNode is used to trace the mobility of nodes

4.5.3 LinkLayer LL

The class LL simulates the link layer, inheriting from class LinkDelay.

4.5.4 ARP

The class ARP simulates the ARP procedure, inheriting from class LinkDelay, the same as the link layer.

Class State

ARPEnter_List arthead_: the list of all known arp entries (the mapping from protocol address to mac address);
MobileNode* node_: the attached mobile node;
Mac* mac_: the mac address of the attached node;

4.5.5 Phy

The class Phy is a pure virtual class needed to be inherited explicitly.

Class State

bandwidth_: the bandwidth of the physical layer.

4.5.6 Network Interface: wirelessPhy

The class WirelessPhy simulates the wireless physical layer, inheriting from the class Phy. Besides working for receiving and sending packets as any other physical layer, it deals with the propagation model (in ns2, mostly 3 models can be used: free space, Two Ray Ground, Shadowing), node sleeping (duty cycle management), energy model management, and maybe different antenna models and modulations. Here, I just describe the transmission behaviors and related features of this class. Any other thing, such as energy, sleeping, is not considered here.

Notice: for the energy consumption tracking simulation, the energy consumption configuration should be done through some commands of this WirelessPhy class, not the energy model class. This is not good implementation of ns2, as I think.

Class States

The bandwidth_ is no more effective here. How to obtain the transmission time through the bandwidth has been moved to the implementation of MAC layer protocol such as MAC802.11. You may see a commented out bind_bw("bandwidth_",&bandwidth_).

Member parameters can be configured through tcl script
double Pt_: the transmitted signal power in Watt;
double freq_: the frequency;
double L_: the system loss factor (mostly 1.0);
double RXThresh_: the receiving power threshold in Watt;
double CStresh_: the carrier sense threshold in Watt;
double CPThresh_: the capture threshold in Watt;
double lambda_: the wavelength in meters. It is calculated through lambda_=SPEED_OF_LIGHT/freq_;
Channel Status: is one of the 4 statuses, sleep, idle, rcv, send, but only sleep, idle are actually used;

other composed member objects:

Antenna* ant_: antenna

Propagation* propagation_: propagation model;

Modulation* modulation_: modulation scheme;

4.5.7 MAC

The class Mac simulates the mac layer object, working as parent class for all kind of mac types;

Class States

The composed objects

LL* ll_: link layer object, up-target;

Phy* netif_: network interface, down-target;

Channel* channel_: down-target of down-target;

Handler* callback_: whose down-target is *this*(mostly, the queue);

MacState state_: deciding the state of the mac or channel, such as idle, recv, send, coll, rts,

cts, ack, polling;

double bandwidth_: the really effective bandwidth of the transmission simulation;

double delay_: the MAC layer computing overhead

MAC Functionality

Since it is just an abstract class, most the functions of it would be overridden by inherited classes (see the following MAC802.11).

4.5.8 IEEE 802.11 MAC

Class States

PHY_MIB phymib_: physical layer management information base (MIB) such as the minimal and maximal size of contention window (CWMin, CWMax), the slot time for each slot in contention window (SlotTime), the SIFS, DIFS, PIFS, EIFS, etc.

MAC_MIB macmib_: mac layer MIB such as the threshold for packet size over which the RTS/CTS would be adapted (RTSThreshold), STA short or long retry limit (ShortRetryLimit LongRetryLimit), failure counters, etc.

bss_id_: for network of infrastructured model;

basicRate_: transmission rate for control packets such as RTS, CTS, ACK and Broadcast;

dataRate_: transmission rate for mac layer data packets;

mhNav_: NAV (network allocating vector) counting down timer;

mhRecv_: receiving timer;

mhSend_: sending timer;

mhDefer_: defer timer;

mhBackoff_: backoff timer;

double nav_: NAV in seconds;

rx_state_: receiving or incoming state (MAC RECV or MAC IDLE);

tx_state_: sending or outgoing state

5

tx_active_: transmitter is active or not;

Packet* pktRTS_: outgoing RTS packet when sending RTS;

Packet* pktCTRL_: outgoing control packet (CTS or ACK);

Packet* pktTx_: the packet needed to be sent out such as data packet or any other packet from upper layer;

cw_: current size of contention window;

ssrc_: current STA short retry counter;

slrc_: current STA long retry counter;

4.6 A New 802.11 implementation and our plan

The existing 802.11 implementation in ns-2.32 doesn't support infrastructure mode simulations. Beacon frame, Scanning, Authentication and Association functions have not been implemented. [3]

Besides the lack of support of infrastructure mode, the original ns2 simulator does not include the access point class, it only include a base station type which is not necessarily an access point.

We base our simulation on the Ilango's ns2 patch [3], where the following functions are have been added to ns2:

1 – Beacon Frames and Passive Scanning

2 – Probe frames and Active Scanning

3 – Authentication

4 – Association

5 – Inter-AP communication

In addition As to Dec 2007, the latest version of Ilango's patch does not support multi channel communication between neighbor BSS and Inter-AP communication exclusive channels.

Based on Ilango's patch, we are going to develop the following functions:

1 – Possible solutions to the AP buffer problem.

2 – Inter-AP communication separate channels and multi-channel BSS support, which will solve the AP buffer problem and probably improve the performance of inter-AP forwarding function.

5. Conclusion and Future work

In this paper we first investigate the IEEE 802.11 infrastructure mode, and point out a critical weakness of the infrastructure mode, which is, the AP buffer could be overflowed by intensive intra-cell connectionless traffic. We demonstrate that this is an inevitable problem with the 802.11 infrastructure mode. We then propose several possible ways to solve this problem. In order to build our own protocol enhancement to 802.11, we do simulations with ns2, a study on basic architecture of ns2 and its wireless extension is presented. Finally an

implementation of 802.11 infrastructure mode is introduced and we give our own plan at last.

Our future work includes the verified solutions to the AP buffer problem, simulation results statistics, and building the Inter-AP communication schemes according to our own design. Further, this design may find its application in wireless mesh network.

6. References

- [1] <http://www.isi.edu/nsnam/ns/>
The Network Simulator - ns-2
- [2] The CMU Monarch Project
<http://www.monarch.cs.cmu.edu/> Computer Science Department, Carnegie Mellon University 1999.
- [3] Ilango Purushothaman, Sumit Roy, "Infrastructure mode support for IEEE 802.11 implementation in NS-2", *report*, University of Washington, Dec 2007.
- [4] IEEE 802.11 part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications, August 1999.
- [5] IEEE 802.11 part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications, Amendment 4: Further higher data rate extension in the 2.4GHz Band, June 2003.
- [6] IEEE 802.11 part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications, Amendment 8: Medium Access Control (MAC) Quality of Service Enhancements, November 2005.
- [7] W. R. Stevens, B. Fenner, and A. M. Rudoff. Unix network programming, the sockets networking API, Vol. 1, 3rd Edition. Addison-Wesley, 2004.
- [8] F. Cali, M. Conti, and E. Gregori. IEEE 802.11 wireless LAN: Capacity analysis and protocol enhancement. In *Proc. IEEE Infocom'98*, pages 142-149, 1998.
- [9] G. Xylomenos and G. Polyzos. TCP and UDP performance over a wireless LAN. In *Proc. IEEE Infocom'99*, pages 439-446, 1999.
- [10] G. Bianchi. Performance analysis of the IEEE 802.11 distributed coordination function. *IEEE Journal on Selected Areas in Communications*, 18(3):535-547, March 2000.
- [11] F. Cali, M. Conti, and E. Gregori. Dynamic tuning of the IEEE 802.11 protocol to achieve a theoretical throughput limit. *IEEE/ACM Transactions on Networking*, 8(6):785-799, Dec. 2000.
- [12] A. Vasani and U. Shankar. An empirical characterization of instantaneous throughput in 802.11b WLANs. Technical Report CS-TR-4389, UMIACS-TR-2002-69, Department of Computer Science and UMIACS, University of Maryland College Park, 2002.
- [13] H. Wu, Y. Peng, K. Long, S. Cheng, and J. Ma. Performance of reliable transport protocol over IEEE 802.11 wireless LAN: analysis and enhancement. In *Proc. IEEE Infocom'02*, pages 599-607, 2002.
- [14] S. Pilosof, R. Ramjee, D. Raz, Y. Shavitt, and P. Sinha. Understanding TCP fairness over wireless LAN. In *Proc. IEEE Infocom'03*, pages 863-872, 2003.
- [15] A. L. Wijesinha, Y. Song, M. Krishnan, V. Mathur, J. Ahn, and V. Shyamasundar. Throughput measurement for UDP traffic in an IEEE 802.11g WLAN. In *Proc. IEEE SNPD/SAWN'05*, pages 220-225, May, 2005.
- [16] S. Choi, K. Park, and C.-K. Kim. On the performance characteristics of WLANs: revisited. In *Proc. ACM Sigmetrics'05*, Banff, Alberta, Canada, June, 2005.
- [17] V. Bychkovsky, B. Hull, A. Miu, H. Balakrishnan, and S. Madden. A measurement study of vehicular internet access using in situ Wi-Fi networks. In *Proc. ACM Mobicom'06*, Los Angeles, CA, USA, September, 2006.
- [18] The UCB Daedalus Group. The daedalus project home page. Available from <http://http://daedalus.cs.berkeley.edu/>, 1998.
- [19] Understanding the implementation of IEEE MAC 802.11 standard in NS-2, Ke Liu, SUNY-Binghamton University.