Estimate the Number of Relevant Images in Infinite Databases Using Two-Order Markov Chain

Xiaoling Wang

Abstract

In image retrieval the most commonly used performance measures are precision and recall. However, to determine the number of relevant images in an infinite database presents a significant challenge as the relevant parameters are not directly observable. In our research, we use internet as a vehicle to investigate this problem, and evaluate search results from major Web Image Search Engines (ISEs). We also investigate whether the cumulative relevance of images in different results pages follows particular stochastic behaviors, such as Two-Order Markov chain whose probability of event occurrence depends on not only a current state but also the immediate past state transitions. From such model, we shall estimate the total number of relevant images for major image search engines.

1 Introduction and Related Work

Due to the increased importance of the Internet, the use of image search engines such as Google, Yahoo, and Msn is becoming increasingly widespread. However, for many web ISEs, it is difficult for users to make a decision as to which web ISE should be selected. It is obvious that the more effective the system is, the more it will offer satisfaction to the users. Therefore, retrieval effectiveness [1], [2], [3], [4] becomes one of the most important parameters to measure the performance of web image retrieval systems [5], [6], [7], [8], [9], [10], [11], [12]. As we know, the most commonly used performance measures are the precision P and recall R [10], [13], [14], [15], [16], [17], [18], [19], [20], [21], but to compute recall R is rather difficult as the total number of relevant images is not directly observable in such a potentially infinite database.

Many researchers have conducted studies to evaluate the retrieval effectiveness of web ISEs. Ece Çakır et al. [7] described the retrieval effectiveness of image search engines based on various query topics. Fuat Uluç et al. [14] described the impact of the number of query words on image search engines. However, none of these studies describe how to estimate the total number of relevant images for the image search engines. All of them only view the first two page results. In the study by Sprink and Jansen [16], data collected from Dogpile was analyzed and one of the findings was that the percentages of the users that viewed only the first page and those that viewed only the first two pages of document search results were about 71% and 15.8%, respectively. Although many works used recall as the measure to evaluate the image search engines, not many papers work on the estimation of the number of relevant images in infinite databases. An algorithm called sample-resample is presented in by Si and Callan [15]; in environments containing resource descriptions already created by query-based sampling, the sample-resample method uses several additional queries to provide an estimate of the database size. Therefore, if the database size has been known, then the distribution of relevant images can be estimated.

In our paper, we model the probabilistic behavior of the distribution of relevant images among the returned results by evaluating the performance of some widespread web ISEs.

In next section, we introduce how to apply Two-Order Markov Chain Model [23], [24], [25], [26] to image retrieval and describe the queries selection. The experimental results and the validation of the models will be discussed in section 3. Finally we summarize our works, and present some directions of future work in the last section.

2 Basic Model and Queries Selection

2.1 Two-Order Markov Chain Model

Since in internet image search, results are returned in units of pages, we shall focus on the integer-valued stochastic process $X_1, X_2, \ldots$, where X_J represents the aggregate relevance of all the images in page J, which may be estimated by

$$X_J = \sum_{i \in J} Z_{Ji} \tag{1}$$

where $Z_{Ji} = 1$ if the i^{th} image on page J is relevant, and $Z_{Ji} = 0$ if the i^{th} image on page J is not relevant.

As to this stochastic modeling of cumulative page image relevance, we shall investigate in particular the Two-Order Markov Chain Model whose probability of event occurrence depends on not only a current state but also the immediate past state transitions. Therefore, a Markov chain of order m where m is finite, is a process satisfying

$$\begin{aligned} &Pr\{X_n = x_n | X_{n-1} = x_{n-1}, \cdots, X_1 = x_1\} \\ = &Pr\{X_n = x_n | X_{n-1} = x_{n-1}, \cdots, X_{n-m} = x_{n-m}\} \end{aligned} \quad (2)$$

Where m is less than n.
In other words, the future state depends on the past m states. It is possible to construct a chain (Y_n) from (X_n) which has the 'classical' Markov property as follows:
Let $Y_n = (X_n, X_{n-1}, \ldots, X_{n-m+1})$, the ordered m-tuple of X values. Then Y_n is a Markov chain with state space S^m and has the classical Markov property.

In our paper, we estimate the relevant images for web ISEs based on Two-Order Markov chain. Therefore, according to what have described above, Y_n is a special sequence, represented as
$Y_n = X_n X_{n-1}, n = 1, 2, 3, \ldots,$
Where Y_n is a Markov chain with a state space S^2 and has the classical Markov property.
The transition probability is a process satisfying

$$\begin{aligned} &Pr\{Y_n = y_n | Y_{n-1} = y_{n-1}\} \\ = &Pr\{X_n X_{n-1} = x_n x_{n-1} | X_{n-1} X_{n-2} = x_{n-1} x_{n-2}\} \end{aligned} \quad (3)$$

And the transition probability is satisfying the condition

$$\begin{aligned} &Pr\{X_n X_{n-1} = x_n x_{n-1} | X_{n-1} X_{n-2} = x_{n-1} x_{n-2}\} \\ &= \begin{cases} 0, & \text{if } x_{n-1} \neq x_{n-1} \\ \geq 0, & \text{if } x_{n-1} = x_{n-1} \end{cases} \end{aligned} \quad (4)$$

Based on our previous study, the state space for One-Order Markov chain model is S={0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20}. The state space for the Two-Order Markov chain should be S^m corresponding, where m is equal to 2. Therefore, the original state vector is a 1-by-441 matrix and the transition probability matrix is a 441-by-441 matrix. We can also effectively estimate the initial probabilities if the sample is large enough. The probabilities are placed in a vector of state probabilities:

$$\pi(J-1, J) = (\pi_{00}, \pi_{01}, \pi_{02}, \pi_{03}, \cdots, \pi_{ij}), \quad (5)$$

Where π_{ij} is the probability having i relevant images in page J and having j relevant images in page $J-1$.
Therefore, from this model, we can estimate the number of relevant images by pages by using the formula in page J:

$$\pi(J-1, J) = \pi(J-1, J-2) * P, J = 2, 3, \ldots, n. \quad (6)$$

Then according to the original state vector and transition probability matrix, we could estimate the total number of relevant images for web ISEs.

2.2 Testing Image Search Engine and Queries Selection

We choose Google, Yahoo and Msn as the testing ISEs. Because from [22], the total market share of Google, Yahoo and Msn are 90.2%, namely, Google [27], Yahoo [28], and Msn [29], whose market shares are 64%, 16.3%, and 9.9%, respectively.
Google (`www.google.com`)
Yahoo (`www.yahoo.com`)
Msn (`www.bing.com`) (new msn image search engine)

In our experiment, total 100 queries are used, which 70% queries are provided by the authors and 30% of them are the popular search suggestion term observed from Google when you submit a keyword. 70% queries consists of one-word, two-word and more than three-word queries, which range from simple words like apple to more specific query like apple computers and finally progressing to rather specific

Table 1: Part of Sample Query list

Categories	**Sample Queries**
One-word Queries	Apple Dolphin Octopus Facebook Roxy Wildlife Skiing Alleyway Maldives Puppy
Two-word Queries	Apple Computer Plane Crash Octopus Card Outer Space Night Scene Daisy Flower Street-Art Baby Cry Afghan Child Twin Towers
Three-word Queries	Man Wearing Hat Macro Fly Eyes Sunrise and Sunset Jordan Basketball Nike Black and White Portrait HongKong Night Scene Flowing in the Wind Michael Schumacher Ferrari Chinese Opera Mask Victoria Harbour HongKong

Table 2: Test Query List

Categories	Test Queries
One-word Queries	Bangkok Stockholm Parkour
Two-word Queries	Sahara Desert Solar System Liu Xiang Abstract Smoke
Three-word Queries	Clown Fish and Sea Anemone Couple Silhouette at Sunset Mexico City Skyline

search, such as the query may be contain three or more elements. While 30% queries consists of the popular suggestion term with the lowest returned results and the largest returned results when we submit a keyword what the Google shows us. A part of selected queries are given in Table 1 and Table 2 shows us the testing queries.

3 Experimental Results

In our experiment, 100 queries are submitted to each of ISE, and we record the number of relevant images page by page. Based on Spink and Jansen's study [12], evaluating the images in first two pages is enough. Such a finding seems useful for the users who only want to find less than forty images. However, it could not satisfy the users' need who wants to search much more images. Therefore, the number of relevant images in the first ten pages is recorded for each query and we will model the Two-Order Markov Chain Model to help us to estimate the number of relevant images per page based on the recorded data. In the following section, we will discuss the results by applying the Two-Order Markov chain model we mentioned in section 2.

3.1 Two-Order Markov Chain Model

Fig. 1 illustrates the actual results returned by test ISE for all the test queries and the estimating results by using Two-Order Markov Chain Model we obtain from the example queries for Google. According to this figure, we can see that the Two-Order Markov Chain Model fits the test queries well, no matter for the one-word query, two-word query or

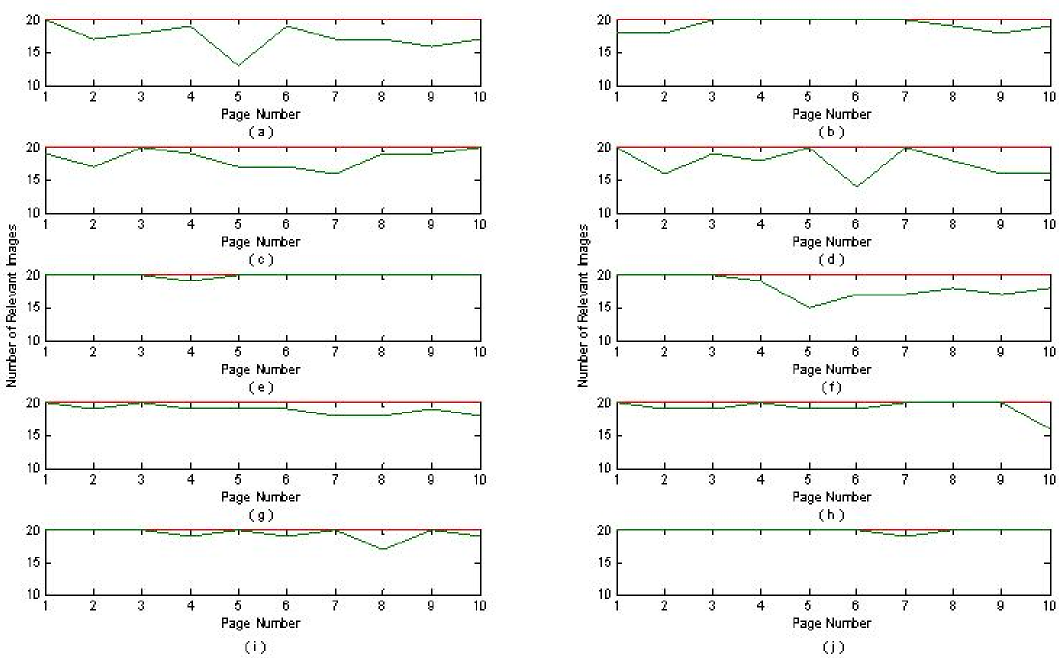

Figure 1: Test results of different queries for Google (a)bangkok. (b)stockholm. (c) parkour. (d) sahara desert. (e) solar system. (f) liu xiang. (g) abstract smoke. (h) clown fish and sea anemone. (i)couple silhouette at sunset. (j)mexico city skyline

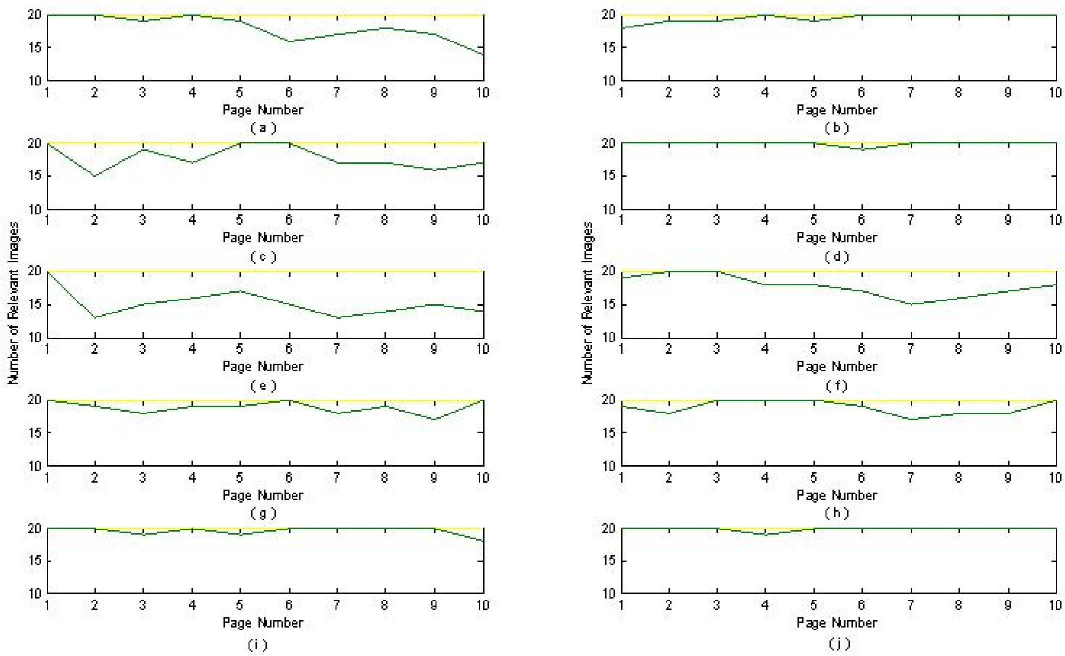

Figure 2: Test results of different queries for Yahoo (a)bangkok. (b)stockholm. (c) parkour. (d) sahara desert. (e) solar system. (f) liu xiang. (g) abstract smoke. (h) clown fish and sea anemone. (i)couple silhouette at sunset. (j)mexico city skyline

three-word query. Although there are some distinctions, the distinctions are not large except for several points. But apparently, the actual results are smaller than the predicting results. However, all of this are based on human judgement, in order to be more precise, we will use MAE to measure whether we can use such a model to estimate the total number of relevant images for the web ISE.

Fig. 2 provides the actual results of different test query for ISE Yahoo. The Figure shows that the Two-Order Markov Chain Model also fits the test queries quite good.Especially for the query called Stockholm, the model fits the data perfectly. And as the results we observed from Google, the actual results are smaller than the predicting results too. But apparently for query Solar System, the model fits the actual results quite bad. However, for most test queries, it is a good predicting model based on human judgment.

Fig. 3 gives the test results of different test query and the predict results given by Two-Order Markov Chain Model for ISE Msn. The Figure tells us that the Two-Order Markov Chain Model fits the test queries also well for the test queries. Meanwhile, it appear that the trend between the actual results and predict results seems quite the same. But for the query couple silhouette at sunset, it seems that we can't use such a model to predict the number of relevant images. Therefore, the model that we investigated is not a perfect model which can perfectly suitable for every query. Finally, we will also use MAE to measure whether we can use such a model to estimate the total number of relevant images for the web ISE.

3.2 Measure of Forecast Accuracy

The mean absolute error (MAE) is a common measure of forecast error in time series analysis. MAE is a quantity used to measure how close forecasts or predictions are to the eventual outcomes. In the previous section, we just judge the accuracy of the models we investigated based on human judgment; however, human judgment is not very precise and well persuade. Therefore, in this section, we will use MAE to measure the Two-Order Markov Chain Model to judge whether the model we investigated can well estimate the number of relevant images per page for major ISEs.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}|e_i| \tag{7}$$

As we know, the smaller the MAE the better the model is. In our paper, that means the smaller the MAE, the better

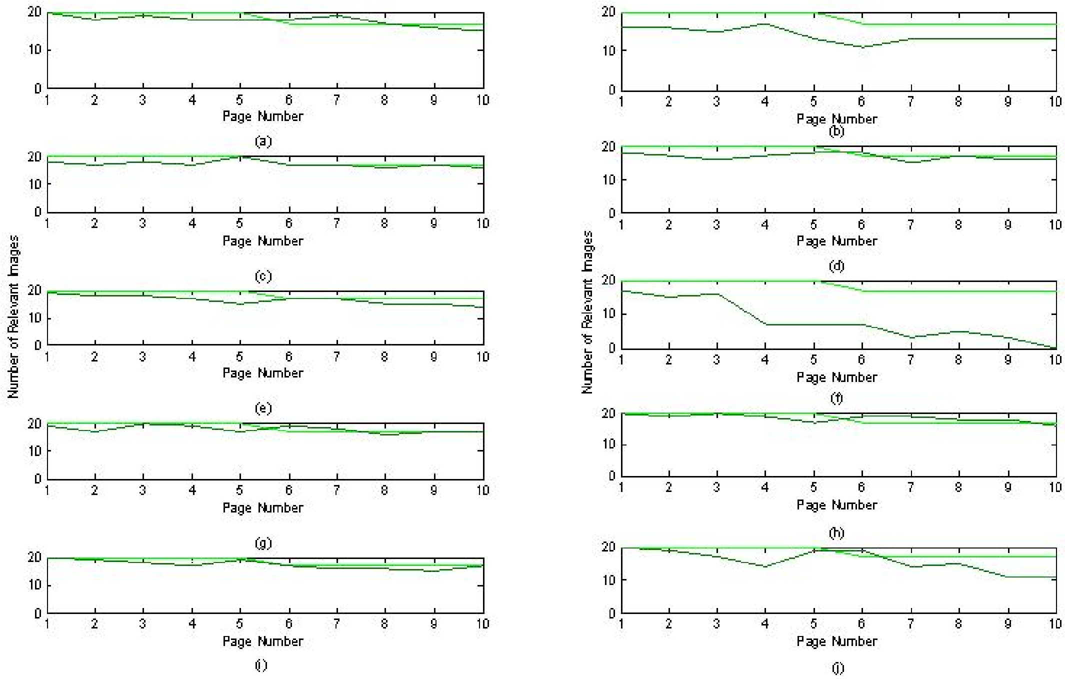

Figure 3: Test results of different queries for Msn (a)bangkok. (b)stockholm. (c) parkour. (d) sahara desert. (e) solar system. (f) liu xiang. (g) abstract smoke. (h) clown fish and sea anemone. (i)couple silhouette at sunset. (j)mexico city skyline

Table 3: MAE for Different Models and Different Search Engines

MAE		Two-Order MC MODEL		
		Google	Yahoo	Msn
One-word Queries	Bangkok	2.7	2.0	1.3
	Stockholm	2.3	0.1	1.9
	Parkour	1.1	1.1	1.2
Two-word Queries	Sahara Desert	0.8	0.5	4.5
	Solar System	0.1	4.8	2.0
	Liu Xiang	0.8	1.1	1.2
	Abstract Smoke	0.1	0.1	3.0
More than three-word Queries	Clown Fish and Sea Anemone	1.7	2.2	1.2
	Couple Silhouette at Sunset	1.9	2.2	10.5
	Mexico City Skyline	0.6	0.4	1.1
MMAE		1.21	1.45	2.79

the model can well estimate the number of relevant images for the web ISEs. The table 3 shows us the results of measurement of forecast accuracy for the model.

According to Table 3, for the Two-Order MC Model, 73.3 percent of MAE is smaller than 2.0, which means the percentage of most deviation between the actual results and the forecast results is less than 10 percent. Therefore, we could conclude that the Two-Order MC Model is good for us to use them to estimate the number of the relevant images for web ISEs. Meanwhile, according to the MAE, we could say that we can use the Two-Order MC Model to estimate the number of relevant images for the web ISEs, because all the MAE listed in the Table 3 is small enough. But there are exceptions; the MAE of query Couple Silhouette at Sunset is 10.5 which are quite large.

4 Conclusion and Future Work

Currently, estimating the number of relevant images in the infinite image search engines is quite hard, but it is so important for us. Therefore, we develop a set of image queries to investigate models to estimate the number of relevant images in infinite ISEs. And using some queries to validate the model we obtain.

In this paper we applied the Two-Order Markov Chain to investigate the probabilistic behavior of the distribution of relevant images among the returned results for the major ISEs. We have carried out experiments based on the returned results of 100 training queries and 10 test queries. For Two-Order Markov Chain Distribution, we were able to construct the transition probability matrix and the original state probability. Then apply the Two-Order Markov Chain to calculate the probability of all probable number of relevant images. Finally we picked up the largest probability and regarded its corresponding number as the number of relevant images of current page. We found that Two-Order Markov Chain Model could well present the distribution of relevant images among the returned results for the major ISEs and the percentage of deviation is less than 10 percent.

Future work includes considering whether the cumulative relevance of images in different results pages follows hidden Markov chain.

References

[1] P. Over, C.H.C. Leung, H. Ip, and M. Grubinger, "Multimedia Retrieval Benchmarks," *IEEE Multimedia,* vol. 11, no. 80-84, Apr. 2004

[2] Nikhil V. Shirahatti, Kobus Barnard, "Evaluating Image Retrieval," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05),* vol. 1, pp. 955-961, 2005

[3] J.R. Smith, "Image Retrieval Evaluation," *IEEE Workshop on Content - Based Access of Image and Video Libraries (CBAIVL),* pp. 112, 1998

[4] Marchand Maillet, S. and Worring, M. "Benchmarking image and video retrieval: an overview." In *Proceedings of the 8th ACM international Workshop on Multimedia information Retrieval (MIR '06),* pp. 297-300, 2006

[5] Jaroslav Pokorný, "Web Searching and Information Retrieval," *Computing in Science and Engineering,* vol. 6, no. 4, pp. 43-48, July/Aug, 2004

[6] K. Stevenson and C.H.C. Leung, "Comparative Evaluation of Web Image Search Engines for Multimedia Applications", In *Proceedings of IEEE International Conference on Multimedia and Expo,* July 2005

[7] Ece Çakır, Hüseyin Bahçeci, Yıltan Bitirim, "An Evaluation of Major Image Search Engines on Various Query Topics." In *Proceedings of the Third International Conference on Internet Monitoring and Protection (ICIMP '08),* pp. 161-165, 2008.

[8] Münevver Tuğçe Elagöz , Mehtap Mendeli , Remziye Zeden Manioğluları , Yıltan Bitirim, "An Empirical Evaluation on Meta-Image Search Engines," In *Proceedings of the Third International Conference on Digital Telecommunications (ICDT '08),* pp. 135-139, 2008.

[9] Tumer, D., Shah, M.A., and Bitirim, Y. , "An Empirical Evaluation on Semantic Search Performance of Keyword-Based and Semantic Search Engines: Google, Yahoo, Msn and Hakia." In *Proceedings of the 2009 Fourth international Conference on internet Monitoring and Protection (ICIMP '09),* pp. 51-55, 2009.

[10] Ishioka, T., "Evaluation of Criteria for Information Retrieval." In *Proceedings of the 2003 IEEE/WIC international Conference on Web intelligence,IEEE Computer Society,* pp. 425, 2003

[11] Jiang, H., "Study on the Performance Measure of Information Retrieval Models." In *Proceedings of the 2009 international Symposium on intelligent Ubiquitous Computing and Education (IUCE '09),* pp. 436-439, 2009.

[12] Demirci, R.G., Kişmir, V., and Bitirim, Y., "An Evaluation of Popular Search Engines on Finding Turkish Documents." In *Proceedings of the Second international Conference on internet and Web Applications and Services (ICIW '07),* pp. 61, 2007

[13] Raquel Kolitski Stasiu, Carlos A. Heuser, and Robertoda Silva, "Estimating Recall and Precision for Vague Queries in Databases", In *Proceedings of Springer-Verlag Berlin Heidelberg(CAISE '05),* LNCS 3520, pp. 187-200, 2005

[14] F. Uluç, E. Emirzade and Y. Bitirim, "The Impact of Number of Query Words on Image Search Engines", In *Proceedings of Second International Conference on*

Internet and Web Applications and Services (ICIW '07), pp. 50, 2007

[15] Si, L. and Callan, J., "Relevant document distribution estimation method for resource selection." In *Proceedings of the 26th Annual international ACM SIGIR Conference on Research and Development in informaion Retrieval (SIGIR '03),* pp. 298-305, 2003

[16] A. Sprink and B.J. Jansen, "Searching Multimedia Federated Content Web Collections," Online Information Review, vol.30, no.5, pp. 485-495, 2006

[17] Jin, X. and French, J.C., "Improving image retrieval effectiveness via multiple queries." In *Proceedings of the 1st ACM international Workshop on Multimedia Databases (MMDB '03),* pp. 86-93, 2003

[18] Black Jr, J.A., Fahmy, G., and Panchanathan, S., "A Method for Evaluating the Performance of Content-Based Image Retrieval Systems." In *Proceedings of the Fifth IEEE Southwest Symposium on Image Analysis and interpretation (SSIAI '02),* pp. 96, 2002

[19] Nottelmann, H. and Fuhr, N., "Evaluating different methods of estimating retrieval quality for resource selection." In *Proceedings of the 26th Annual international ACM SIGIR Conference on Research and Development in informaion Retrieval (SIGIR '03),* pp. 290-297, 2003

[20] Hauptmann, A.G. and Christel, M.G., "Successful approaches in the TREC video retrieval evaluations." In *Proceedings of the 12th Annual ACM international Conference on Multimedia (MULTIMEDIA '04),* pp. 668-675, 2004

[21] Longzhuang Li, Yi Shang, "A new statistical method for performance evaluation of search engines." In *Proceedings of the 12th IEEE international Conference on Tools with Artificial intelligence (ICTAI '00),* pp. 208, 2000

[22] Nielsen NetRatings, "Google accounts for more than half of all web searches, while Yahoo and Msn combined account more than one fourth, according to Nielson Netrating", April 2009. http://www.polepositionmarketing.com/emp/april-2009-search-engine-2/

[23] Thomas W. Miller, *Data and Text Mining: A Business Applications Approach.* Upper Saddle River, N.J.: Pearson Prentice Hall, c2005

[24] W.S. Kendall, F. Liang, J.S. Wang, *Markov chain Monte Carlo: innovations and applications,* Singapore; Hackensack, NJ: World Scientific, c2005

[25] Render, R., R.M. Stair Jr., and M.E. Hanna. *Quantitative Analysis for Management-Chapter 16,* 10th Edition, Prentice Hall, 2009

[26] Dani Gamerman, *Markov chain Monte Carlo: stochastic simulation for Bayesian inference,*London; New York : Chapman and Hall, 1997

[27] Google Search Engine `www.google.com`

[28] Yahoo Search Engine `www.yahoo.com`

[29] Msn Search Engine `www.bing.com`