Soc. Netw. Anal. Min.
DOI 10.1007/s13278-011-0023-y

ORIGINAL ARTICLE

Social opinion mining for supporting buyers' complex decision making: exploratory user study and algorithm comparison

Li Chen · Luole Qi

Received: 3 November 2010 / Revised: 21 February 2011 / Accepted: 15 March 2011

Abstract This article reports our study of the role of social content (i.e., user-generated content in social networking environment) in online consumers' decision process when they search for an inexperienced product to buy. Through close observation of users' objective behavior and interview of their reflective thoughts during an initial exploratory user study, we have first derived a set of system implications and integrated these implications into a three-stage system architecture. Furthermore, driven by the specific implication regarding the impact of user reviews in influencing users' decision stages, we have presented a linear-chain conditional random-field-based social-opinion-mining algorithm, and have identified its higher effectiveness against related algorithms in an experiment. Finally, we present our system's user interfaces and emphasize on how to display the opinion-mining results in the form of both quantitative presentation and qualitative visualization.

Keywords Users' information needs · Social content · Complex decision making · Inexperienced products · Decision system · Opinion mining

1 Introduction

With the advance of social networking techniques and applications, an increasing number of users have come to this platform to behave not only as visitors, but also as contributors. As a result, a huge amount of user-generated content has appeared, such as user reviews given to products and user photos shared in popular media sites such as Flickr. It is hence crucial to investigate how such content can be exploited to enhance current decision support systems, such as e-commerce services, so that online buyers can benefit from the utilization of other users' contributed content in making a better and confident purchase decision.

As a matter of fact, social content has always been recognized to play an important role in a consumer's hybrid decision process, in which the decision maker seeks advices with the aim of reducing uncertainty (Cialdini and Goldstein 2004; Kim and Srivastava 2007; Lee et al. 2006). In recent years, intelligent decision supports have been developed for assisting users in efficiently processing various types of product information and enabling them to make more informed and accurate decisions. Some studies have also been done to analyze user logs and transaction data in the Web sites so as to guide e-commerce promotions (Adnan et al. 2011; Raeder and Chawla 2011). Among these approaches, the so-called *recommender system* is a typical example, as it emphasizes on the value of user-generated product content (e.g., ratings/reviews) to determine whether some items would be preferred by a user given that her neighbors rated these items high (Adomavicius and Tuzhilin 2005; Siersdorfer and Sizov 2009).

However, this kind of decision system has been mainly oriented to low-value, and frequently purchased products such as books, movies or music, for which the current user can provide ratings or prior history. They are limited and difficult to be applied to expensive, infrequently purchased products (such as digital cameras, computers and cars). The reason is that for these products, it is difficult to infer the current user's interests/preferences as s/he would

L. Chen (✉) · L. Qi
Department of Computer Science, Hong Kong Baptist University, Hong Kong, China
e-mail: lichen@comp.hkbu.edu.hk

L. Qi
e-mail: llqi@comp.hkbu.edu.hk

Published online: 14 April 2011

not have prior purchase experience. A simple recommendation method, as popularly applied in current sites, is hence to rely on the user's real-time viewing behavior to return a set of related items (e.g., "people viewed this product also view others…" or "people bought this product also bought…"). Unfortunately, little effort has been made to enhance this recommendation process by incorporating other kinds of social content.

Another limitation in the related areas is that little attention has been paid to understanding users' needs on *opinion features*. For instance, besides knowing their criterion on the optical zoom for a camera, it should be also interesting to see whether they are concerned about "ease of use" as commented by other consumers. The reason for calling such feature as *opinion feature* is because its values can be only derived from other users' reviews. In the previous decade, a number of studies were conducted to extract features and their associated opinions from user reviews, which is termed as "opinion mining". However, though some algorithms have been proposed including rule-based (Turney 2002), statistic methods (Hu and Liu 2004) and recent hidden Markov model (HMMs)-based approach (Jin et al. 2009), they are still limited in identifying hidden and non-independent features and opinions.

1.1 Our contributions

Thus, driven by the above limitations, in this paper, we have engaged in achieving two objectives as highlighted in the following.

First, we were interested in understanding deeply users' information needs when they were involved in the search for an inexperienced product. This understanding can help us identify the exact role of social content in consumers' decision process and hence show insights into how improvements can be made on existing decision supports and how the opinion-mining technique can be incorporated into it. Indeed, our study was grounded on Adaptive Decision Theory (Payne et al. 1993), which states that in complex decision environment, users are usually uncertain about their targets at the start, but construct their product preferences in an adaptive, constructive nature. Moreover, users' decision behavior generally follows two stages (Häubl and Trifts 2000): they will first screen down the number of available alternatives to a reduced consideration set, and then they will in-depth examine the selected candidates to make the final choice. To us, the question is, at each stage, what kind(s) of social content would be most relevant to the decision maker, given that various user-generated resources can be obtained now from social networking media (e.g., Epinions and Flickr). Through an exploratory user study, we first obtained answers to these concerns. Furthermore, based on this study's results, we were able to suggest a refined three-stage system architecture. The corresponding system implications motivated us to invest more effort in building an effective social-opinion-mining method.

Thus, our second contribution in this article was that we experimentally compared two typical supervised learning models for mining social opinions: lexical hidden Markov models (L-HMMs) and linear-chain conditional random fields (CRFs). Since the latter approach naturally consider arbitrary, non-independent features without conditional independence assumption, we in detail describe how opinion-mining process can be accomplished through it, and how it could outperform the L-HMMs-based method in extracting various types of feature entities and opinion polarities.

Finally, we introduce how the opinion-mining outcomes are presented in our user interfaces. Two alternatives are realized: one is a quantitative presentation by means of normalizing subjective opinion values into numerical scores, and the second is a qualitative visualization of opinion words via the format of tag cloud. These interfaces are targeted to provide a visualized and interactive display so as to facilitate users to make more informed and effective decisions.

The article is hence organized as follows: we first survey related works on decision systems and opinion-mining algorithms (Sect. 2). In Sect. 3, we present our research questions and the proposed methods. The exploratory user study then follows (Sect. 4), and we describe the study's setup, materials used, participants recruited, procedure and analysis of statistical results. In Sect. 5, we give the set of practical implications from the user study and then point out the motivation for follow-up studies (Sect. 6). Section 6 emphasizes on the opinion-mining algorithm and compares CRFs-based method with L-HMMs-based one, followed by user interface designs. Section 7 concludes our work, indicates its limitations and future directions.

2 Related work

2.1 Related work on decision supports for inexperienced products

Researchers from marketing refer consumer behavior as the action and decision process of people who purchase goods and services (Engel et al. 1990; Foxall et al. 1998). The action starts with the need of a product or a service that arises in the customer's mind and then goes through the process of information searching and product evaluation to lead to purchase decision and post-purchase evaluation. For different product categories, consumer buying behavior will differ. That is, people make their decision differently when they are involved in buying a high-value product

such as a camera compared to the situation of buying a low-value product such as a book (Engel et al. 1990; Olshavky and Granbois 1979). The amount of cognitive effort applied to the decision-making process is in nature directly related to the level of importance that the consumers place on purchase of the specific product. As for complex ones that are expensive and infrequently experienced (e.g., cameras, computers, cars), extensive decision-making effort is commonly applied by consumers in seeking information and deciding.

Accordingly, as mentioned before, researchers from the psychology field described a basic two-stage process in such complex decision environment, where the depth of cognitive load and information processing varies (Payne et al. 1993). Based on the two-stage model, Häubl and Trifts (2000) demonstrated that recommendation agent (RA) is more useful for the initial screening to increase the quality of consideration set, and the use of comparison matrix (CM) is more effective in facilitating pair-wise product comparisons at the second stage to improve objective decision quality. Some researchers have also investigated the impact of other factors. For instance, Knijnenburg et al. proved that adjusting the elicitation of users' multi-attribute preferences to their domain knowledge can significantly augment individual satisfaction with the system (Knijnenburg and Willemsen 2009). Al-Qaed and Sutcliffe (2006) have proposed an adaptive decision support system architecture (ADSS) aimed at providing information display, searching strategies and appropriate advice based on a set of pre-defined decision rules, to adapt to the consumer's present product domain. Example-based systems, including FindMe (Burke et al. 1997), Dynamic-Critiquing (McCarthy et al. 2005) and Example-Critiquing (Pu and Chen 2006; Chen and Pu 2006), have been also proposed to support users to give improving feedback to example products, in the form of critiques (e.g., "I would like something cheaper", "with faster processor speed"). Mahmood and Ricci (2007) modeled such conversational process as a sequential decision problem based on the Markov decision process (MDP), which involves different user states and actions.

In the last decade, recommender systems have also been widely developed to aid users in making right choices, by suggesting some items that they might not have found by themselves. However, as mentioned in the introduction, most systems were targeted at frequently experienced public taste products (e.g., music, movie, Web page, book, events) (Adomavicius and Tuzhilin 2005; Groh and Ehmig 2007; Kayaalp et al. 2011). For example, the classical user–user collaborative filtering technique is with the assumption that the current user can provide ratings on a set of items, as they have experienced them before (Groh and Ehmig 2007). This approach is hence unlikely to be applied to high-value, inexperienced products. Recent extension on recommender systems still focus on low-value product domains, though they have started to exploit other types of social resources, such as tags, friendship and membership, to compensate for the limitation of pure rating-based methods (Siersdorfer and Sizov 2009; Yuan et al. 2009; Guy et al. 2010).

Thus, though it has been claimed that shoppers tend to wait for early adopters' opinions to reduce the risk of buying a new product (Cialdini and Goldstein 2004), few studies have in depth explored users' social information needs for inexperienced product search. Indeed, in this area, most researches have been still grounded on products' static attributes to establish users' preference model. For instance, in Example-Critiquing (Pu and Chen 2006), a multi-criteria user model is built via eliciting users' criteria on multiple basic, static attributes. Most e-commerce applications usually display the social content for users to browse, and few incorporate them into the process of decision support and recommendation generation. Leino and Räihä (2007) have conducted a tentative experiment that measured product ratings and reviews as part of recommendations in influencing users' searching strategies. However, because it is a preliminary study, it is still not clear about what social information users require and how they process them across their decision stages.

2.2 Related work on opinion mining

In another research field that is called *opinion mining* (or *sentimental classification*), many researchers have attempted to adopt natural language processing techniques and data mining tools to extract meaningful opinion values from product reviews (called documents in some literature) (Dave et al. 2002; Pang and Lee 2008). Opinion mining has been conducted either at the document level or at the feature level. At the document level (which is to obtain an overall opinion value for the whole document), Turney (2002) used point-wise mutual information (PMI) to calculate an average semantic orientation score of extracted phrases for determining the document's polarity. Pang et al. (2002) examined the effectiveness of applying machine-learning techniques to address the sentiment classification problem for movie review data. Hatzivassiloglou and Wiebe (2000) studied the effect of dynamic adjectives, semantically oriented adjectives and gradable adjectives on a simple subjectivity classifier, and proposed a trainable method that statistically combines two indicators of gradability. Wilson et al. (2005) proposed a system called OpinionFinder that automatically identifies when opinions, sentiments, speculations and other private states are present in text, via the subjectivity analysis. Das and Chen (2001) studied sentimental classification for financial documents. However, although the above studies are all related to sentiment classification, they use sentiment to

represent a reviewer's overall opinion and do not find which features the reviewer actually liked and disliked. For example, an overall negative sentiment about an object does not mean that the reviewer dislikes every aspect of the object. It can only indicate that the average opinion as summarized from the review is negative.

To discover in-depth a reviewer's opinions on almost every aspect that s/he mentioned in the text, some researchers have tried to mine and extract opinions at the feature level. Hu and Liu (2004) proposed a feature-based opinion summarization system that captures highly frequently featured words by using association rules under a statistical framework. It extracts the features of a product that customers have expressed their opinions on and concludes with an opinion score for each frequent feature, while ignoring infrequent features. Popescu and Etzioni (2005) improved Hu and Liu's work by removing frequent noun phrases that may not be real features. Their method can identify part of a relationship and achieve better precision, but shows a small drop in recall. Scaffidi et al. (2008) presented a new search system called Red Opal that examined prior customer reviews, identified product features and then scored each product on each feature. These scores were used to determine which products to be returned when a user specifies a desired product feature.

However, because the above studies are mostly unsupervised learning mechanisms, it is unavoidable that their accuracy is limited and many other types of entities (such as component, function, dependent features, etc.) cannot be identified through them. Thus, some researchers have attempted to adopt more precise supervised learning models to increase the opinion-mining efficacy. One typical work is OpinionMiner (Jin et al. 2009). It was built based on lexicalized hidden Markov model (HMMs) which can integrate multiple important linguistic features into an automatic learning process. However, its limitation is that it cannot represent distributed hidden states and complex interactions among labels. It can neither involve rich, overlapping feature sets. That is why in our work we have employed another model, conditional random field (CRFs) (Lafferty et al. 2001), because it naturally considers arbitrary, non-independent features without conditional independence assumption. Although lately some investigators, including Miao et al. (2010), have also attempted to adopt CRFs to perform sentiment analysis, they did not use CRFs to identify feature-based polarity orientation or use it to extract various types of feature entities.

3 Research questions and our methods

From the related works, we can see that there are two major branches of studies related to social opinions: one is at the system level, but few of the related studies have incorporated social opinions to enhance user decisions for inexperienced products; and another is at the algorithm level, but opinion-mining algorithms' results have been less studied regarding their merit in supporting consumer decisions. Our goal was then to build a bridge between them. In this section, we first list a set of research questions that were not well addressed in related studies, and then give our research methods to answer these questions.

3.1 Research questions

Our first objective was to identify, at the system level, the importance of social content relative to static product attributes in users' purchase decision process, when they search for inexperienced products. Specifically, through the empirical method of studying consumers' behavior, we expect to answer the following questions:

Question 1: do consumers in reality follow the general two-stage decision process, or can a more precise decision model be identified?

Question 2: at each stage, how does social content (e.g., product reviews) practically act to assist the user in processing information and making decisions?

Question 3: could we develop more effective decision systems to optimize the merits of social content and adapt them to users' actual decision needs?

Answers to the above questions can enable us to build a system infrastructure and, in the system, we can in particular explore user reviews at the algorithm level. The following objective is then to study social-opinion-mining algorithms. More questions for this objective are:

Question 4: would CRFs-based opinion-mining approach perform better than other learning models in terms of extracting feature entities and social opinion values?

Question 5: how could the opinion-mining results be informatively represented on user interfaces for supporting users to make effective decisions?

3.2 Research methods

3.2.1 Exploratory user study

To answer the first three questions, we conducted an exploratory user study with the aim of understanding users' natural decision behavior. The specific aims were: (1) to trace users' decision-making process so as to refine the basic two-stage model; and (2) to understand users' social information needs at each stage. Thus, we on purpose asked users to use existing online commercial sites that provide plentiful and diverse product information, so that we could observe users' actual information needs. Our

primarily studied social content was *user reviews* (in the form of natural languages), as shared by consumers based on their post-purchase evaluation experiences. Besides, we also attempted to involve other types of social content so as to reveal user reviews' relative value. Among various resources, popularity information (e.g., "top products") is typical, which is based on the statistics of consumers' actual usages or purchases. "Related products" (e.g., "people viewed this product also viewed others") is also a popular social type, because it correlates the user's current view/click with other consumers' clicking or purchase behavior.

In Sect. 4, we will in detail describe how the user study was set up and analyzed, including materials used, participants recruited, procedure and results.

3.2.2 System design and algorithm development

The follow-up work was an extension to the exploratory study. Concretely, the user study results suggest not only a three-stage decision process model, but also a set of system implications that we could adopt for improving current systems. To optimally utilize and present social content, we subsequently emphasize on feature extraction and social opinion mining. For this part, we investigated how to extract more accurate opinion features from user reviews and how to present them in an informative and interactive way to users. Concretely, with respect to the algorithm contribution, we applied the linear-chain CRFs-based learning model to mine opinions and proved the algorithm's outperforming efficacy by comparing it with other supervised learning approaches. From the perspective of user interface, we proposed a visualization method based on the tag cloud format and provided multiple interaction functions for facilitating users to easily examine the opinion-mining results.

4 User study setup

4.1 Materials

Thus, in this experiment, our goal was to record users' decision behavior when they were confronted with various information resources. We first classified all kinds of product-related information into two principal categories: *static features* that include all in-born attribute values about the product (e.g., the digital camera's price, weight, megapixels, optical zoom, etc.), and *social features* defined as any sort of data that require other consumers' contribution (e.g., product reviews, product popularity, etc.).

We used Flickr Camera Finder as one experiment material, since it provides popularity data (e.g., "most popular cameras", "trends in brands", "trends in camera use", etc.) according to statistics of the Flickr community members who had uploaded images with a particular camera over a certain time. Such *usage-driven popularity generation* is inherently different from traditional e-commerce sites' purchase or promotion-based popularity. In addition, this site also supplies other sorts of social content such as the product's usage trend analysis.

A standard e-commerce site was also offered in the study, which mainly provides user reviews a complete set of static product features, traditional popularity information (so as to be compared with Flickr's), "related products", etc. To choose this site, we investigated a number of options, e.g., Amazon, Yahoo Shopping, shopping.com, etc., and finally selected Yahoo Shopping because it not only uses the same product database as Flickr Camera Finder (CF), but also can be representative of other sites regarding amount of information, diversity and presentation.

The experiment was then designed in a free-choice scenario where users were allowed to freely select and examine any product information that can be obtained from the two sites: Flickr Camera Finder (http://www.flickr.com/cameras/, *Flickr CF* for short) and Yahoo Shopping (shopping.yahoo.com) for cameras (*Yahoo CF* for short). It is worth noting that our experiment was not a comparative user study and our purpose was not to evaluate the Web sites. As indicated above, our objective was to reveal which social contents are relevant to buyers' needs and how the content could be best exploited to generate more intelligent decision supports.

4.2 Participants

Each participant was required to take at least 2 h in the experiment. One hour was spent in performing a user task by freely using the two sites and another hour was to answer a set of interview questions. We finally recruited 12 volunteers (three females) who were all motivated to take the study because they were interested in buying a digital camera at the time of our experiment. They were master's or PhD students in our university (ages ranging from 20 to 40 years). Besides, all of our studied subjects made their first-time encounters to the two Web sites (i.e., Flickr CF and Yahoo CF), and so their behavior was not biased by any prior usage. Regarding their familiarity with the camera's static features, most of them (nine users) were moderately familiar or had little familiarity.

We also asked some general questions about their online shopping experiences before the formal study started. All users indicated that they had the experience. Many of them (9 users) had bought items at least every 3 months in e-stores, while most of the frequently purchased items were of relatively low values such as books, accessories, DVDs.

On being asked how they purchased a high-value product, the majority (10) replied that they examined as much product information as possible to find the best one. The examined product information was mostly from online media, especially ones that provide consumer reviews, but they seldom bought the product online (due to concern about delivery or security). It can be hence seen that the online environment has been at least adopted as an information-seeking platform for consumers to construct product preferences, and other users-generated contents seemed to play an important role in absorbing them to this platform.

4.3 Procedure

The user task was: "Imagine you are prepared to buy a digital camera. Please use the assigned sites to examine product information that you care, and identify the product that meets your needs". There was an administrator present in each user study to control the whole session. At the beginning, an initial warm-up period (10 min) was given to each participant for her/him to be familiar with the two sites' facilities as much as possible, so that when the task formally started, her/his actions was mainly driven by her/his actual information needs. During the formal trial, their interaction actions, including on-screen mouse moves, clicks and keyboard inputs, were all automatically captured by a screen observer software (i.e., Morae). After the participant accomplished the task, a semi-structured interview was conducted by the administrator to get her/his reflective thoughts.

4.4 Result 1: users' decision process

Question 1: do consumers in reality follow the general two-stage decision process, or a more precise decision model can be identified?

To answer the first research question, we analyzed all subjects' natural decision behavior. The results surprisingly showed that they all exhibited a precise three-stage decision process: (1) to screen all alternatives and select ones for in-depth evaluation; (2) to view the product's details and save it in the wish list if it is near satisfactory; (3) to compare candidates in the wish list and make the final choice.

Moreover, the transition between these three stages is not in the supposed sequential order, but is iterative in nature and the size of the consideration set gradually decreases. Figure 1 (left) gives an example of how the three-stage process was conducted by a representative user.

Concretely, at the start, all users were with some initial preferences in mind. As they elaborated during the post-study interview, the preferences were mostly need oriented, e.g., looking for a camera that is "easy to use", "easy to carry", "with colorful images", "of high cost performance", "better for night scenes" or "better for long distance picture-taking". Several users (6) also had some criteria on static features (e.g., on price, type, megapixels, screen size, battery or focal length).

Further analysis of their interaction logs showed that they all considered no more than four brands within their whole decision session. In fact, they all first narrowed down to the most preferred brand and sought its alternatives' basic information (Stage 1, see Fig. 1, right). If any product(s) interested them when they browsed the brand's product list, they went on to examine the product's details; they saved it in their wish list if it was near satisfactory (Stage 2). After examining one brand, they switched on to another preferred brand and performed a similar process. This iterative cycle between Stages 1 and 2 continued until a set of candidates was determined (see detailed analysis later). At this point, they entered into Stage 3 to compare candidates in their wish list and confirmed the final choice.

Due to the common behavior discovered in all participants, we came up with a three-stage decision process model (see Fig. 1, right). Each stage was with its input and output. Comparing the three-stage to the originally suggested two-stage model by Payne et al. (1993), we found

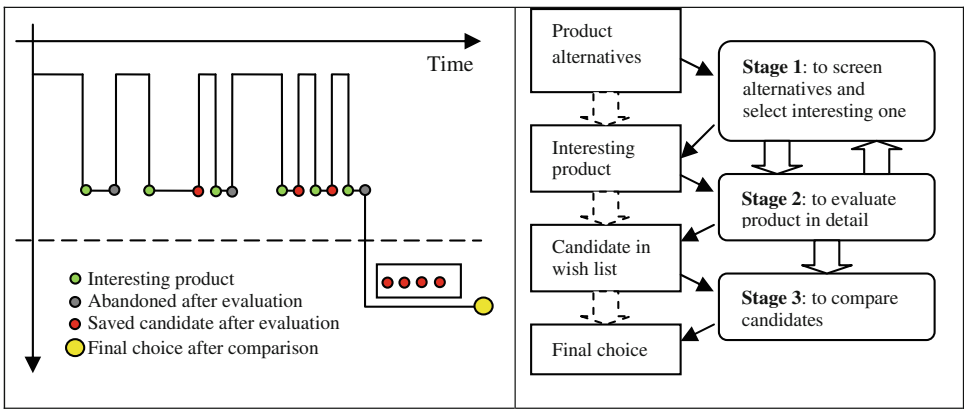

Fig. 1 Example of a user's searching behavior and the three-stage consumer decision process

that it indeed provides a more elaborate view. Previous concerns such as how users narrow down to a consideration set from a range of available alternatives (i.e., the first stage in the two-stage model) were well clarified in our study and detailed into two stages.

4.5 Result 2: information processing at each stage

After identifying the three decision process stages, we were then engaged in obtaining an answer to the second question:

Question 2: at each stage, how does social content (e.g., user reviews) practically act to assist the user in processing information and making decisions?

4.5.1 Stage 1: when screening out interesting products

4.5.1.1 Users' objective behavior At this stage, by replaying the video log that recorded each user's interactive actions, we found out how many products were selected (i.e., clicked by the user to see details) and from where these products were located. It indicated that on average, 9.67 (SD = 4.78) products were chosen to view details, among which 5.42 were located in Yahoo and 4.25 were in Flickr CF. Figure 2 (left) concretely shows the distribution of these products' locations. Specifically, basic static features (as provided by Yahoo for browsing and filtering) provided the highest chances that enabled the average user to obtain 39.79% interesting products. The second and third winners were Flickr CF's popularity-based sorting list (27.51%) and brand popular list (12.18%), respectively. In comparison, Yahoo's popularity list got much less hits (5.28%). There were only two participants who accessed "Top Digital Cameras" in Yahoo, against nine who consulted the popularity ranking in Flickr CF.

The remaining products were located either from the results of keywords search (e.g., the user inputted a pre-known model for search) (6.53%), or coincidently discovered through Flickr's image-related products (4.83%) or Yahoo's related products (i.e., "shopper who viewed this also viewed …", 3.89%).

Thus, in total, above half of the picked products (53.69%) originated from social contents (see Fig. 2, right). In particular, product popularity was shown to be more active than others for users to identify interesting ones at the first stage. As one user said, "popularity is a suitable proxy to measure the product's quality when I am not familiar with a brand or uncertain about what I want". It was also regarded as "the best form of recommendations" in this condition.

4.5.1.2 Users' qualitative comments Users' qualitative comments obtained during the post-study interview further exposed their reflective thoughts particularly about product popularity and "Related Products", when they have processed them at the first stage.

Credibility of product popularity When being asked why they went to Flickr CF for accessing "product popularity", when the similar kind of info was also available in Yahoo CF, most of them responded that because it was perceived more trustworthy in Flickr CF: "I trust the information on the social forum"; "I trust Flickr's popularity information because of its large amount of users"; "Flickr is more neutral and credible"; "Although this is my first-time using this website, the information sounds credible since it should be based on actual usages." They felt that product popularity based on community's image uploading statistics is surprising at the first impression. They were soon used to it and inclined to refer to this popularity ranking whenever needed. The product popularity on Yahoo Shopping site (e.g., "top Digital Cameras"), however, was perceived "less trustworthy", because "The 'top products' in Yahoo may be only dependent on users' clicks or for companies' promotion purpose." "The popularity information in Yahoo may be faked. It looks more trustworthy

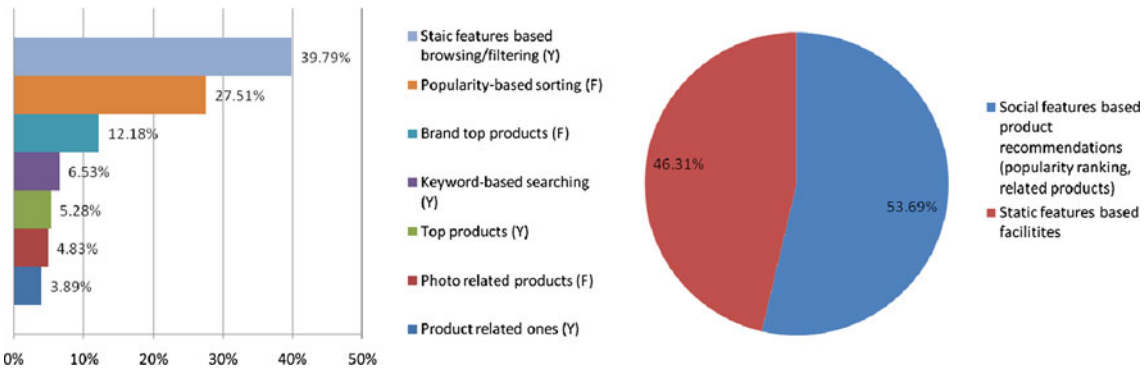

Fig. 2 *Stage 1* Locations of products which were picked for in-depth evaluation (*Y* for Yahoo CF, *F* for Flickr CF), and their overall distribution over social and static sources

and real in Flickr." "Flickr is more neutral because it is a consumer-operated website. The information on Yahoo may be not so real because it is more commercial-oriented." It can be hence inferred that users have the propensity to trust the data from the social media site like Flickr, because it is perceived more dependent on a large community's real usages and less of commercial interests.

Additionally, users suggested several ways to improve the generation of product popularity. For instance, one user suggested that Flickr community members' geographical distribution should be considered, since "one camera model was suitable for European, but probably not for Chinese", "people from the same cultural background may have similar preferences". Another user proposed to involve the time property. He commented that "it should be easier to compare the popularity values of different products if they were released at the same time."

"Related Products" to be integrated with expert opinions Relative to product popularity, "Related Products" (i.e., "shoppers viewed this product also viewed …") were less referred. Users commented that if these recommendations could integrate experts' professional opinions about the relevance a product to others, they would look more meaningful than being purely dependent on other consumers' clicking behavior. As one user said, "imagine the friends around you all use Canon, you would be not familiar with Nikon. But if there is a comparison table from an expert explaining what their differences are, I will go to see Nikon's products." Another user also noted "because people sometimes just randomly clicked, the information from 'shopper viewed this product also viewed others' cannot be so credible. Experts' suggestions can be more useful to be regarded as important references." Some users further suggested that the recommendations could be even better if they take into account of the users' hard constraints (e.g., on price range, product type), because "The 'related products' are useful, but I will not be interested in them if they are out of my price expectation." "If the products are with the type that I prefer, I will more likely consider them." Thus, users' comments can explain why they were not so active in adopting "Related Products" when selecting products to examine. It also infers that if this type of recommendations could be well integrated with experts' opinions and also be matched to the user's hard attribute constraints, their adoption chance could be potentially increased.

4.5.2 Stage 2: when evaluating a product in detail

4.5.2.1 Users' objective behavior At the second stage, we were interested in investigating what detailed info that users evaluated after they picked a product from the Stage 1. The analysis of their page visits indicates that 42.86% of products were evaluated on Yahoo (that provides the product's user reviews, full specifications), 30.44% on Flickr CF (that provides the product's usage trend analysis and community images), and 26.70% on both sites' products' detail pages. Among all of evaluated products, 45.82% were put into the average user's wish list (i.e., mean = 4 products, SD = 1.95). The page evaluations respectively contributed 39.09% (1.50 products), 6.25% (0.25) and 91.67% (2.17), to establishing her wish list (the % means the percent of products saved as candidates among these with the same type of evaluation, see Fig. 3). It hence infers that the examination of product details from both Flickr CF and Yahoo CF can most likely convince the user to take the product as a candidate. The correlation is indeed highly significant ($p < 0.001$) by Pearson coefficient. Another fact is that 91.7% users' final choices were products that underwent this combined review.

4.5.2.2 Users' qualitative comments *User reviews: negative versus positive ones* At this stage, we particularly discovered the role of user reviews. Most users stated that the reviews are very important information for them to evaluate a product's details, since "they help me judge the product's true quality". They said that they were always motivated to see the detailed review info if the product rating was low. They liked the separation of reviews into *pros* and *cons* categories since it eases their comparison. More notably, negative reviews were more useful than positive ones, because: "the motivation of buying a product is not because it is very perfect, but is whether you can stand its drawbacks"; "Every product should have flaws, and what I want to get from user reviews is whether they can disclose these negative aspects". All participants agreed that "I will not buy a product only because it has positive ratings and reviews, but will certainly not buy it if it has negative reviews, especially on features that I am concerned about". Moreover, some users commented that the number of user reviews also takes effect: "few reviews will have low credibility", but it is still better than zero since "in the case that two products both have few user reviews, I will still read the reviews to get the feeling of which product would be better."

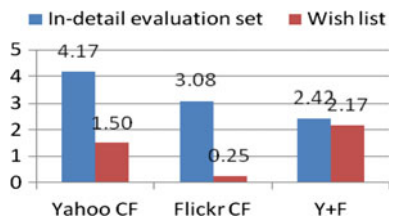

Fig. 3 *Stage 2* Locations of products which were evaluated in detail and the amounts of products which were saved as candidates in the average user's wish list

4.5.3 *Stage 3: when comparing candidates and confirming the final choice*

At the last stage when users nearly came to making the "purchase" decision, they all conducted a comparison across candidates in their wish lists. In order to know which factors they considered at this point, we recorded items they have viewed after their wish list was established (i.e., when no new product was added in the list). It shows that 66.7% (8 out of 12) users went to Flickr CF to compare candidates' usage trends or images as contributed by Flickr community, and 33.3% concentrated on user reviews or specifications as given in Yahoo CF (see Fig. 4, left).

Figure 4 (right) shows that totally 75% of our participants in fact relied on social contents, including usage trends, product reviews, and community images, against 25% users who focused on static features. Therefore, social features are demonstrated more influential at this stage. Users' qualitative comments also reflected that "I would like to rely on the social content to identify which product should be better than others"; "The product's usage trend can help me form a correct judgment and reduce the uncertainty from purely evaluating its static specifications."

5 Practical implications from the user study

Thus, through the exploratory user study, we in-depth studied users' natural decision behavior and their reflective thoughts. The study not only identified how a three-stage decision process was practically conducted (see Sect. 4.4), but also revealed what and how social contents were processed by users when they were engaged in different stages (Sect. 4.5). These two findings well answered our research questions 1 and 2. For the next step, given these findings and users' qualitative comments, we were able to derive a set of practical implications for the system development (to question 3):

Question 3: could we develop more effective decision systems to optimize the merits of social content and make them adapt to users' actual decision needs?

5.1 System implications for Stage 1

As for decision Stage 1 "screening out interesting products", two types of social contents were found relevant: product popularity and "related products". Regarding product popularity, it is suggested that it should better originate from social media sites (like from Flickr), because in such platforms, it can reflect a community of like-minded users' real usages, and hence be perceived more credible than in traditional e-commerce sites. Moreover, the popularity info is more referential and helpful when users have not formed their clear target at the start. This implies that it could well complement standard feature-based browsing facilities, and adapt to different users' preference-certainty levels. To provide a better support, it was also revealed that the popularity can be further customized to involve contextual factors, such as regional and time properties, so as to be dynamically matching to the current user's context.

"Related Products" can be also likely enhanced by integrating with expert opinions and involving users' stated feature constraints. By these ways, these recommendations could be more interesting to the user and hence be with higher adoption chance.

5.2 System implications for Stages 2 and 3

The role of user reviews was mainly discovered at Stages 2 and 3, i.e., when users evaluate a product in detail, and when users compare all candidates before making the final choice. As shown before, most of users intentionally relied on them to judge a product's quality. Moreover, users indicated that they cared more about negative reviews than positive ones, which is in accordance with a pervious claim

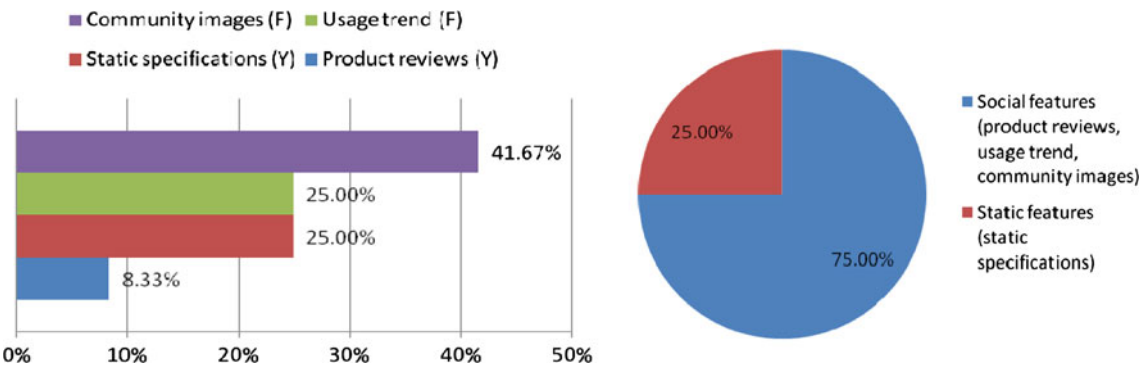

Fig. 4 *Stage 3* Factors that % users mainly considered during the product comparison before they confirmed the final choice, and the overall distribution over social and static sources

that "when information about an object or firm comes through the opinions of another person, negative information can be more credible and generalizable than positive information" (Mizerski 1982). We therefore believe that the mining and exposure of opinion values and the indication of their relevance to user needs can likely support users to conduct more accurate product evaluation and final decision.

The implication is also applied to improving comparison facility at Stage 3. The traditional support as broadly appearing in current e-commerce sites, is *comparison matrix*, by which users can put multiple products in an *alternatives* (rows) × *attributes* (columns) matrix. This support has been demonstrated to allow for higher decision quality than the condition without it (Häubl and Trifts 2000). However, most of existing applications are limited to only display static features in the matrix. Given our finding that around 75% users tended to consult with social features during the comparison, we believe that embedding products' social values in the matrix could be much helpful to improve users' decision quality.

Figure 5 summarizes these system implications in the three-stage architecture.

6 Follow-up work: social opinion mining

Driven by above implications, especially ones about the role of user reviews in facilitating users' product evaluation and comparison, in our follow-up work, we have focused on the algorithm development and user interface design. Concretely, we have targeted to produce feature-level automatic sentimental classification of user reviews and extract opinion features. As a result, we expect that a recommendation set of {*feature*, *opinion value*} pairs, with higher weight placed on negative opinions, can be generated. Such list could be hence not only placed at the product's detail page to supplement standard product specifications, but also be integrated into the comparison matrix to optimize users' comparison performance.

Technically speaking, to address limitations of related opinion-mining techniques (see Sect. 2.2), we have particularly studied the conditional random field (CRFs), which is a discriminative, undirected graphical model that can potentially model the overlapping, dependent features (Lafferty et al. 2001). As mentioned before, although there have been some works of adopting CRFs for document-level segmentation and semantic labeling, few have actually studied its performance in achieving opinion-mining goals. Moreover, previous works did not use CRFs to identify feature-based polarity orientation and not use it to extract various types of feature entities (including infrequent ones).

Therefore, we have developed a CRFs-based opinion-mining algorithm to compensate for related works' limitations. For this part of work, we have three specific questions. (1) How to define feature functions to construct and restrict the linear-chain CRFs model? (2) How to automatically extract different types of product entities and associate appropriate opinion polarities with them? In the following, we will in detail describe how these questions were addressed in our work.

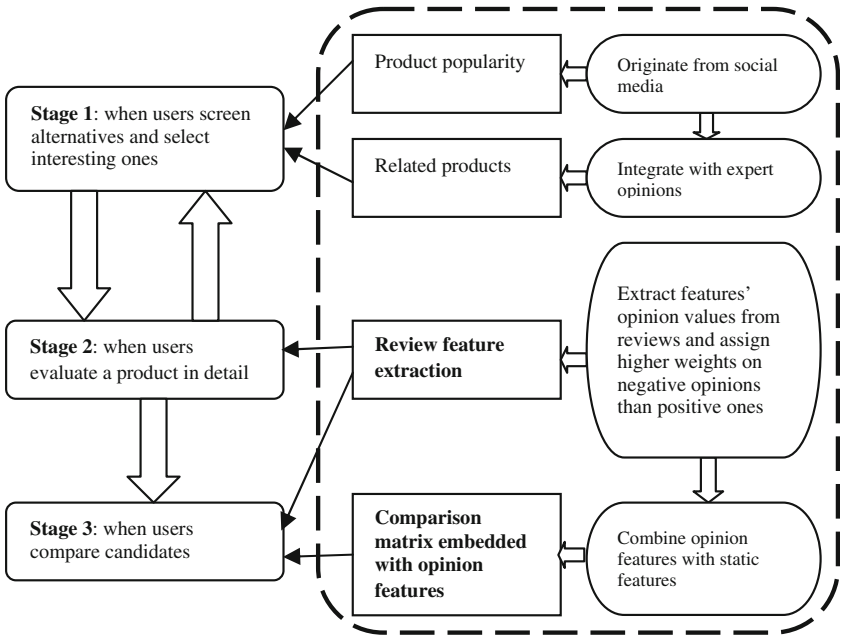

Fig. 5 Users' three-stage decision process and corresponding system implications to support them, as derived from the user study's results

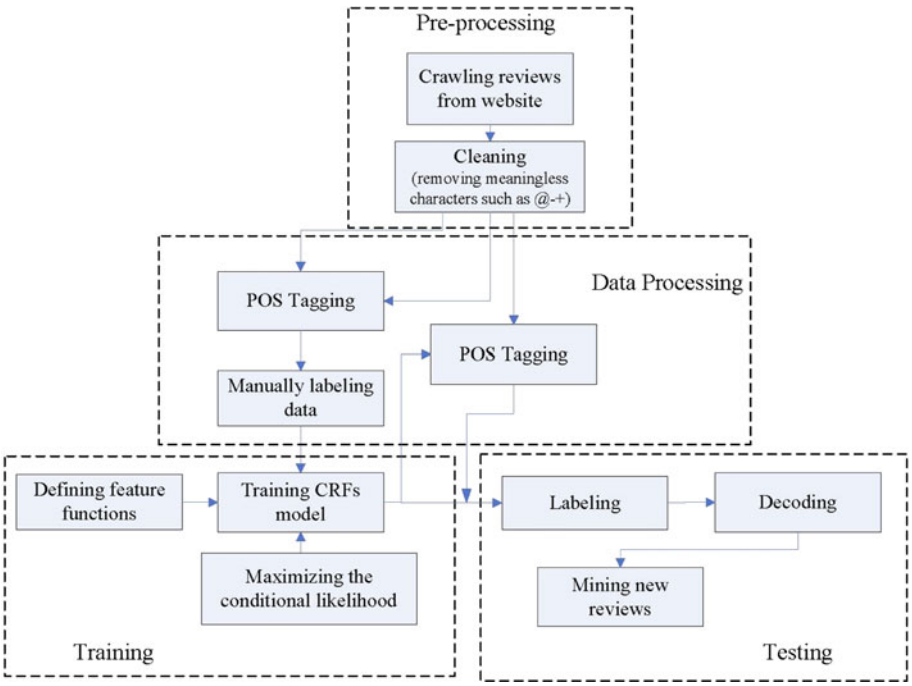

Fig. 6 The control flow of our CRFs-based opinion-mining algorithm

In Fig. 6, we first give an overview to our opinion-mining process. It is divided into four major steps: (1) pre-processing, which includes crawling raw review data and cleaning; (2) tagging data for training the learning model; (3) defining feature functions for CRFs and training it by maximizing the conditional likelihood; and (4) applying the model to label product entities from new review data.

6.1 CRFs learning model

In this section, we first provide some background info about CRFs. Conditional random fields (CRFs) are conditional probability distributions on an undirected graph model (Lafferty et al. 2001; Fei and Fernando 2003; McCallum 2003). It is formally defined as follows: considering a graph $G = (V, E)$ for which V indicates the nodes and E indicates the edges. Let $Y = (Y_v)_{v \in V}$, and (X, Y) is a CRF, where X is the set of variables over the observation sequences to be labeled (e.g., a sequence of textual words that form a sentence), and Y is the set of random variables over the corresponding sequence. The (X, Y) obeys the Markov property with respect to the graph (e.g., part-of-speech tags for the words' sequence). Formally, the model defines $p(y|x)$ which is globally conditioned on the observation of X:

$$p(y|x) = \frac{1}{Z(x)} \prod_{i \in N} \phi_i(y_i, x_i) \tag{1}$$

where $Z(x) = \sum_y \prod_{i \in N} \phi_i(y_i, x_i)$ is a normalization factor over all states for the sequence x. The potentials are normally factorized on a set of features f_k, such as

$$\phi_i(y_i, x_i) = \exp\left(\sum_k \lambda_k f_k(y_i, x_i)\right) \tag{2}$$

Given the model defined in (1), the most probable labeling sequence for an input x is hence:

$$\widehat{Y} = \arg\max_y p(y|x) \tag{3}$$

In our case when we build the model, we take all the nodes V of the graph as states including observed states and hidden states. We use X to represent the observed states and use Y to represent the hidden states. The edges E of the graph are relationships among all the states, which are formally defined by the feature functions (see Sect. 6.3).

6.2 Our problem statement

Based on this model, our goal was then to apply it to practical extract different types of product entities and opinions from textual reviews. According to (Jin et al. 2009), there are four types of entities that inherently exist in a product review: component, function, feature[1] and opinion. Table 1 lists the four types of entities and their examples.

It then came to the question of how we could extract these entities via CRFs. To solve the problem, we first defined three types of tags: entity tag, position tag and opinion tag. We use the category name of a product entity

[1] Please note that the "feature" here refers to the product's feature. It is different from the definition of feature in feature functions (for which we will discuss later when constructing CRFs model).

Table 1 Four types of product entities and their examples, according to (Jin et al. 2009)

Entity	Description and examples
Feature	Properties of components or functions, e.g., color, speed, size, weight
Component	Physical objects of a product, e.g., cell phone's LCD
Function	Capabilities provided by a product, e.g., movie playback, zoom, automatic fill flash, auto focus
Opinion	Ideas and thoughts expressed by reviewers on product, features, components, and functions, e.g., good, satisfying

to be the entity tag. As for a word which is not an entity, we use the character 'B' to represent it. Usually, an entity could be a single word or a phrase. For the phrase-entity, we assign a position to each word in the phrase. Any word of a phrase has three possible positions: the beginning of the phrase, the middle of the phrase and the end of phrase. We use characters 'B', 'M' and 'E' as position tags to respectively indicate the three positions. As for "opinion" entity, we further use characters 'P' and 'N' to respectively represent Positive opinion polarity and Negative opinion polarity, and use "Exp" and "Imp" to respectively indicate explicit opinion and implicit opinion. Here, explicit opinion means that the user expresses opinion in the review explicitly and implicit opinion means the opinion needs to be induced from the review. These tags (i.e., P, N, Exp, Imp) are all called opinion tags. Thus, with above defined tags, we can tag any word and its role in a sentence. For example, the sentence. "The image is good and its ease of use is satisfying" from a camera review is labeled as:

> The(B) image(Feature-B) is(B) good(Opinion-B–P-Exp) and (B) its(B) ease(Feature-B) of(Feature-M) use(Feature-E) is(B) satisfying(Opinion-B–P-Exp).

In this sentence, 'image' and 'ease of use' are both features of the camera and 'ease of use' is a phrase, so we add '-B', '-M' and '-E' to specify the position of each word in the phrase. 'Good' is a positive, explicit opinion expressed on the feature 'image', so its tag is Opinion expressed on the feature image, so its tag is 'Opinion-B–P-Exp' (such tag combination is also called hybrid tags in (Jin et al. 2009). Other words which do not belong to any entity categories are assigned the tag 'B'.

Therefore, if we could get a word's tag, we can know which product entity it refers to and identity the opinion polarity if it is an "opinion" entity. By this way, the task of opining mining can be transformed to an automatic labeling task. The problem can be then formalized as: given a sequence of words $W = w_1 w_2 w_3 \ldots w_N$ and its corresponding parts of speech $S = s_1 s_2 s_3 \ldots s_N$, the objective is to find an appropriate sequence of tags which can maximize the conditional likelihood of (3). The resulting equation is then like this:

$$\widehat{T} = \arg\max_T p(T|W,S) = \arg\max_T \prod_{i=1}^{N} p(t_i|W,S,T^{(-i)}) \tag{4}$$

In (4), $T^{(-i)} = \{t_1 t_2 \ldots t_{i-1} t_{i+1} \ldots t_N\}$ (which are tags in our case, and called hidden states in the general concept). From this equation, we can see that the tag of a word at position i depends on all the words $W = w_{1:N}$, part-of-speech $S = s_{1:N}$ and tags. Unfortunately, it is very hard to compute with this equation as it involves too many parameters. To reduce the complexity, we employ linear-chain CRFs as an approximation to restrict the relationship among tags. In the linear-chain CRF, all the nodes Y in the graph (see Fig. 7) form a linear chain and each feature involves only two consecutive hidden states. Equation (4) can be hence rewritten as:

$$\widehat{T} = \arg\max_T p(T|W,S) = \arg\max_T \prod_{i=1}^{N} p(t_i|W,S,t_{i-1}) \tag{5}$$

6.3 Defining feature functions

From the model above, we can see that in order to make the model more computable, the relationships need to be defined between the observation states $W = w_{1:N}$, $S = s_{1:N}$ and hidden states $T = t_{1:N}$, so as to reduce unnecessary calculations. Thus being the important construct of CRFs, feature functions are crucial to resolve the problem. Let $w_{1:N}$ and $s_{1:N}$ be the observations (i.e., words' sequence and their corresponding parts of speech), $t_{1:N}$, be the hidden labels (i.e., tags). In our case of linear-chain CRFs (see (5)), the general form of a feature function is $f_i(t_{j-1}, t_j, w_{1:N}, s_{1:N}, j)$, which looks at a pair of adjacent states t_{j-1}, t_j, the whole input sequence $w_{1:N}$ and $s_{1:N}$ and the current word's position j. For example, we can define a simple feature function which produces binary value: the returned value is 1 if the current word w_j is "good", the corresponding part-of-speech s_j is "JJ" (which means single adjective word) and the current state t_j is "Opinion":

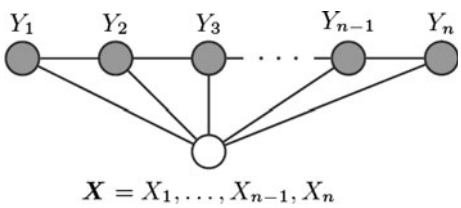

Fig. 7 Linear-chain CRFs' graph structure

$$f_i(t_{j-1}, t_j, w_{1:N}, s_{1:N}, j) = \begin{cases} 1 & \text{if } w_j = \text{good},\ s_j = JJ \text{ and } t_j = \text{Opinion} \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Combining the feature function with (1) and (2) we have:

$$p(t_{1:N}|w_{1:N}, s_{1:N}) = \frac{1}{Z}\exp\left(\sum_{n=1}^{N}\sum_{i=1}^{F}\lambda_i f_i(t_{j-1}, t_j, w_{1:N}, s_{1:N}, j)\right) \tag{7}$$

According to (7), the feature function f_i depends on its corresponding weight λ_i. That is if $\lambda_i > 0$, and f_i is active (i.e., $f_i = 1$), it will increase the probability of the tag sequence $t_{1:N}$, and if $\lambda_i < 0$, and f_i is inactive (i.e., $f_i = 0$), it will decrease the probability of the tag sequence $t_{1:N}$.

Another example of feature function can be like:

$$f_i(t_{j-1}, t_j, w_{1:N}, s_{1:N}, j) = \begin{cases} 1 & \text{if } w_j = \text{good},\ s_{j+1} = \text{NN} \text{ and } t_j = \text{Opinion} \\ 0 & \text{otherwise} \end{cases}. \tag{8}$$

In this case, if the current word is "good" such as in the phrase "good image", the feature functions in (6) and (8) will be both active. This is an example of overlapping features which L-HMMs cannot address.

Specifically, we define several types of feature functions to specify state-transition structures between *W*, *S* and *T*. The different state transitions are based on different Markov orders for different classes of features. For instance, the first-order feature functions are defined as:

1. The assignment of current tag t_j only depends on the current word. The feature function is represented as $f(t_j, w_j)$.
2. The assignment of current tag t_j only depends on the current part of speech. The feature function is represented as $f(t_j, s_j)$.
3. The assignment of current tag t_j only depends on both the current word and the current part of speech. The feature function is represented as $f(t_j, s_j\, w_j)$.

The three types of feature functions are first order, by which the inputs are examined in the context of the current state only. We also define *first-order plus transition* feature functions and second-order feature functions, which are examined in the context of both the current state and previous states. We do not define third order or higher-order feature functions, because they would create data sparsity problem. Table 2 shows the types of feature functions we have defined in our model.

6.4 Training CRFs model

After the graph and feature functions are defined, the model is fixed. The purpose of training is then to identify all the values of $\lambda_{1:N}$. Normally, one may set $\lambda_{1:N}$ according to the domain knowledge, but in our case, we learn $\lambda_{1:N}$ from the training data. The fully labeled review data is $\{(w^{(1)}, s^{(1)}, t^{(1)}),\ldots, (w^{(M)}, s^{(M)}, t^{(M)})\}$, where $w^{(i)} = w^{(i)}_{1:N_i}$ (which is the *i*th words' sequence), $s^{(i)} = s^{(i)}_{1:N_i}$ (i.e., the *i*th part-of-speech sequence), and $t^{(i)} = t^{(i)}_{1:N_i}$ (i.e., the *i*'th tags' sequence). Given that in CRFs we defined the conditional probability $p(t|w, s)$, the aim of parameter learning is to maximize the conditional likelihood with the training data, according to (9):

$$\sum_{j=1}^{m}\log p(t^{(j)}|w^{(j)}, s^{(j)}) \tag{9}$$

To avoid over fitting, the likelihood can be penalized by some prior distributions over the parameters. A commonly used distribution is zero-mean Gaussian: if $\lambda \sim N(0, \sigma^2)$, (9) will become

$$\sum_{j=1}^{m}\log p(t^{(j)}|w^{(j)}, s^{(j)}) - \sum_{i}^{F}\frac{\lambda_i^2}{2\sigma^2} \tag{10}$$

The equation is concave, so λ has a unique set of global optimal values. We learn parameters by computing the gradient of the objective function, and apply the gradient in an optimization algorithm called Limited memory BFGS (L-BFGS).

Formally, the gradient of the objective function is computed as:

$$\begin{aligned} &\frac{\partial}{\partial\lambda_k}\sum_{j=1}^{m}\log p(t^{(j)}|w^{(j)}, s^{(j)}) - \sum_{i}^{F}\frac{\lambda_i^2}{2\sigma^2} \\ &= \frac{\partial}{\partial\lambda_k}\sum_{j=1}^{m}\left(\sum_{n}\sum_{i}\lambda_i f_i(t_{n-1}, t_n, w_{1:N}, s_{1:N}, n) - \log T^{(j)}\right) \\ &\quad - \sum_{i}^{F}\frac{\lambda_i^2}{2\sigma^2} = \sum_{j=1}^{m}\sum_{n} f_k(t_{n-1}, t_n, w_{1:N}, s_{1:N}, n) \\ &\quad - \sum_{j-1}^{m}\sum_{n} E_{t'_{n-1}, t'_n}[f_k(t'_{n-1}, t'_n, w_{1:N}, s_{1:N}, n)] - \frac{\lambda_k}{\sigma^2} \end{aligned} \tag{11}$$

In equation (11), the first term is empirical count of feature *i* in the training data, the second term is the expected count of this feature under the current trained model, and the third term is generated by the prior distribution. Hence, this derivative measures the difference between the empirical count and the expected count of a feature under the current model. Suppose that in the training data, a feature f_k actually appears *A* times, while

Table 2 Defined feature function types and their expressions

Feature function type	Expressions
First order	$f(t_i, w_i), f(t_i, s_i), f(t_i, s_i, w_i)$
First order + transition	$f(t_i, w_i)f(t_i, t_{i-1}), f(t_i, s_i)f(t_i, t_{i-1}), f(t_i, s_i, w_i)f(t_i, t_{i-1})$
Second order	$f(t_i, t_{i-1}, w_i), f(t_i, t_{i-1}, s_i), f(t_i, t_{i-1}, s_i, w_i)$

under the current model, the expected count of f_k is B: when $|A| = |B|$, the derivative is zero. Therefore, the training process is to find λs that can make the two counts equal.

6.5 Algorithm evaluation

Question 4: would CRF-based opinion-mining approach perform better than other learning models in terms of extracting product entities and social opinion values?

In order to answer the fourth research question (we proposed in Sect. 3.1), we conducted an experiment that systematically compared the CRFs-based opinion-mining algorithm with two related methods: L-HMMs-based method (Jin et al. 2009), which is also a supervised model-based learning technique, and rule-based technique as the baseline. The reason that we did not compare with the statistical method from Hu and Liu (2004) was because they treated all product entities (e.g., component, function) as rather than handling them individually.

Concretely, we evaluated these approaches' performance on three metrics: recall, precision and F-measure. Recall is $\frac{|C \cap P|}{C}$ and precision is $\frac{|C \cap P|}{P}$, where C and P are the sets of correct and predicted tags, respectively. F-measure is the harmonic mean of precision and recall, i.e., $\frac{2RP}{R+P}$ (where R is the recall value and P is the precision value). We did not use label accuracy to test each type of tag, because for some tags (e.g., "B" for background words) it is not meaningful to evaluate them. Instead, we emphasize the precision, recall and F score results on the four types of entities, e.g., feature, component, function, and opinion, when comparing the CRFs-based opinion-mining method with the L-HMMs-based one.

We used two datasets of product reviews: one was crawled from Yahoo Shopping, and another was the corpus shared by Hu and Liu (2004). For example, Fig. 8 gives one digital camera's user review (in XML format) that we crawled from Yahoo Shopping site, for which we mainly focused on the "Posting" part which gives the user-generated textual comments.

We finally collected 476 reviews (with 10,769 words) in total from 3 cameras and 1 cell phone, and manually labeled by them by using the 17 distinct tags as defined in

<Review>
<Title>Great Camera</Title>
<Reviewer>l_infante69</Reviewer>
<CreateTime>1133976475</CreateTime>
<HelpfulRecommendations>3</HelpfulRecommendations>
<TotalRecommendations>4</TotalRecommendations>
- <Ratings>
<Rating ratingType="Features">5</Rating>
<Rating ratingType="Overall">5</Rating>
<Rating ratingType="Quality">5</Rating>
<Rating ratingType="Support">5</Rating>
<Rating ratingType="Value">5</Rating>
</Ratings>
<OverallRating>5</OverallRating>
<Pro>Light weight, great battery power</Pro>
<Con>PC Picture Software and Users Guide</Con>
<Posting>This is a great camera. I shopped around and got a great price. This is my first digital camera. No problems with the pictures or the screen. The battery power is fantastic, the size is great, and the pictures and photo options are really nice.

The user guide isn't very user friendly. If you are not electronic savy, it may take some time to figure out this camera.

The software to load the pictures on my PC is also not very user friendly. The only way I can crop and edit pictures is by loading into a different application (such as HP photo director).</Posting>
</Review>

Fig. 8 A digital camera's user review in XML format. The <Posting> … </Posting> part gives the textual comment that we emphasized in the algorithm

Sect. 6.2. After the pre-process of removing meaningless characters (e.g., @-+), we applied the LBJPOS tool[2] to produce the part-of-speech tag for each word. All tagged data were then divided into 4 four sets to perform fourfold cross-validation: one set was used as the validation data for testing and the other three sets were used as training data. The cross-validation process was repeated four times, and every time one set was randomly selected as the testing data. Afterward, the results were averaged to produce the final precision, recall and F-measure scores.

6.5.1 Compared algorithms in the experiment

Rule-based method Motivated by (Turney 2002), we designed a rule-based method as the baseline for comparison. The first step was performing part-of-speech (POS) task. One example of POS result is:

(PRP I) (VBD used) (NNP Olympus) (IN before) (, ,) (VBG comparing) (TO to) (NN canon) (, ,) (PRP it) (VBD was) (DT a) (NN toy) (, ,) (NNP S3) (VBZ IS) (VBZ IS) (RB not) (DT a) (JJ professional) (NN camera) (, ,) (CC but), (RB almost) (VBZ has) (NN everything) (PRP you) (VBP need) (. .)

In the example, each word gets a tag of POS such as NN (noun word), JJ (adjective word), etc. We then applied several basic rules, according to (Hu and Liu 2004; Jin et al. 2009), to extract product features (note: in the rule-based method, components and functions are taken as features).

[2] http://l2r.cs.uiuc.edu/~cogcomp/software.php.

1. One rule is that a single noun that follows an adjective word or consecutive adjective words (such as JJ + NN or JJ) will be regarded as a product entity.
2. Any single noun word that connects an adjective word to a verb will be taken as a product entity, such as NN + VBZ + JJ.
3. Any consecutive noun words that appear at the position described in (1) or (2) will be taken as a product entity phrase.

As for opinion entities, the adjective words that appear in rules 1 and 2 will be, and their sentimental orientation was determined by a lexicon with polarities for over 8,000 adjective words.[3]

L-HMMs method Jin et al. (2009) integrated linguistic features such as part-of-speech results and lexical patterns into a hidden Markov model (HMMs). Their aim was to maximize the conditional probability as defined in:

$$\widehat{T} = \arg\max_T p(W,S|T)p(T) = \arg\max_T p(S|T)p(W|T,S)p(T)$$
$$= \arg\max_T \prod_{i=1}^{N} \left\{ \begin{array}{l} p(s_i|w_1 \ldots w_{i-1}, s_1 \ldots s_{i-1}, t_1 \ldots t_{i-1} t_i) \\ \times p(w_i|w_1 \ldots w_{i-1}, s_1 \ldots s_{i-1} s_i, t_1 \ldots t_{i-1} t_i) \\ \times p(t_i|w_1 \ldots w_{i-1}, s_1 \ldots s_{i-1}, t_1 \ldots t_{i-1}) \end{array} \right\}$$

Three assumptions were made for simplifying the problem: (1) the assignment of the current tag depends not only on its previous tag but also on the previous J words; (2) the appearance of the current word is assumed to depend not only on the current tag, the current POS, but also on the previous K words; (3) the appearance of the current POS depends on both the current tag and previous L words (J = K = L). Then their objective was to maximize.

$$\arg\max_T \prod_{i=1}^{N} \left\{ \begin{array}{l} p(s_i|w_{i-1}, t_i) \\ \times p(w_i|w_{i-1}, s_i, t_i) \\ \times p(t_i|w_{i-1}, t_{i-1}) \end{array} \right\}$$

Maximum likelihood estimation (MLE) was used to estimate the parameters. Other techniques were also used in their approach, including information propagation via entity synonyms, antonyms and related works, and token transformation.

6.5.2 Experiment results and discussion

Table 3 shows the experiment results of recall, precision and *F* score by comparing the three methods, from which we can see that the CRFs-based learning method (henceforth CRFs) increases the accuracy regarding almost all the four types of entities, except the slightly lower recall and *F* score than L-HMMs method (henceforth L-HMMs) in respect of component entity.

Table 3 Experimental results from the comparison of the three approaches: baseline: the rule-based opinion-mining method, L-HMMs: the lexicalized hidden Markov model-based learning method, CRFs: the linear-chain conditional random-field-based learning method

	Rule based	L-HMMs based	CRFs based
Feature entities (%)			
R	–	78.6	81.8
P	–	82.2	93.5
F	–	80.4	87.2
Component entities (%)			
R	–	96.5	91.8
P	–	95.3	98.7
F	–	96.0	95.1
Function entities (%)			
R	–	58.9	80.4
P	–	81.1	83.7
F	–	68.2	82.0
Opinion entities (%)			
R	25.1	53.7	65.3
P	22.5	76.9	84.2
F	23.8	63.2	73.5
All entities (%)			
R	27.2	72.0	79.8
P	24.3	83.9	90.0
F	25.7	77.1	84.3

R recall, *P* precision, *F* *F* score

More specifically, CRFs improve the precision from 83.9 to 90.0% on average and the *F* score from 77.1 to 84.3% in comparison with L-HMMs. Two major reasons can lead to this result. Firstly, L-HMMs assume that each feature is generated independently of hidden processes. That is, only tags can affect each other and the underlying relationships between tags and words/POS-tags are ignored. Secondly, L-HMMs do not model the overlapping features. As for recall, it was also averagely improved from 72.0% by L-HMM to 79.8% by CRF. This is promising because the recall can be likely affected by tagging errors. For example, if a sentence s correct tags should be "Opinion-B–P-Exp, Opinion-M-P-Exp, Opinion-M-P-Exp, Opinion-E-P-Exp" but it was labeled as "Opinion-B–P-Exp, Opinion-E-P-Exp, Opinion-M-P-Exp, Opinion-M-P-Exp", the labeling accuracy is 75%, but recall is 0.

Moreover, besides the feature functions defined in Table 2, we also tested other feature functions in our experiment. Table 4 shows these additional feature functions. Thus, we have seven types of feature functions in total. We assigned each type a number from 1 to 7. We varied their combinations and used them in training

[3] http://www.cs.cornell.edu/People/pabo/movie-review-data/review_polarity.tar.gz.

Table 4 Evaluated feature function types in the experiment

Feature function type	Expressions
1. First order	$f(t_i, w_i)$, $f(t_i, s_i)$, $f(t_i, s_i, w_i)$
2. First order + transition	$f(t_i, w_i)f(t_i, t_{i-1})$, $f(t_i, s_i)f(t_i, t_{i-1})$, $f(t_i, s_i, w_i)f(t_i, t_{i-1})$
3. Second order	$f(t_i, t_{i-1}, w_i)$, $f(t_i, t_{i-1}, s_i)$, $f(t_i, t_{i-1}, s_i, w_i)$
4. First order + transition	$f(t_i, w_i)f(t_i, t_{i-2})$, $f(t_i, s_i)f(t_i, t_{i-2})$, $f(t_i, s_i, w_i)f(t_i, t_{i-2})$
5. First order + transition	$f(t_i, w_{i-1})f(t_i, t_{i-1})$, $f(t_i, s_{i-1})f(t_i, t_{i-1})$, $f(t_i, s_{i-1}, w_{i-1})f(t_i, t_{i-1})$
6. First order + transition	$f(t_i, w_{i-1})f(t_i, t_{i-2})$, $f(t_i, s_{i-1})f(t_i, t_{i-2})$, $f(t_i, s_{i-1}, w_{i-1})f(t_i, t_{i-2})$
7. Second order	$f(t_i, t_{i-2}, w_i)$, $f(t_i, t_{i-2}, s_i)$, $f(t_i, t_{i-2}, s_i, w_i)$

process. Table 5 shows the results from different combinations.

The results show that when we adopted all the feature functions, the precision, recall and *F* score are all better than other varieties. Comparing the condition that used feature functions' combination $1 + 2 + 3 + x$ (x is from 4 to 7) with the condition of $1 + 2 + 3$, we found that the feature functions with number 4 and number 7 can cause higher precision and *F* score, and feature functions with number 5, 6 and 7 can lead to higher recall and *F* score. It hence infers that when more types of feature functions are involved in training CRFs model, it is more likely that the algorithm accuracy can be further improved. The finding motivates us to do more experiments in the future to test more types of feature functions and hence identify the optimal combination.

In the experiment, we also conducted a comparison between the 2 datasets: 238 reviews from (Hu and Liu 2004) and 238 reviews from Yahoo Shopping. The procedure is that we first trained the CRFs on the first dataset and then tested it with the second one, and then did the opposite way. Figure 9 shows the precision, recall and *F* scores averaged over four entities from this comparison. It can be seen that there is no big difference between the two datasets. The largest distance occurs with the precision on feature entity, but it is only 3.9%. The result hence implies that our approach can achieve a stable performance with different datasets and can be hence scalable to mine review data from various resources.

It is also worth noting that the overlapping feature functions as defined in our CRFs model can be useful to discover infrequent entities, which however were often ignored in related approaches. For example, although the entity ISO only appears once in our data, functions $f(t_i, s_i\ w_i)$ and $f(t_i, w_i)$ can be still active in finding this feature. Moreover, the uneasily discovered entities, such as non-noun product entities and non-adjective opinions, can be also identified through our approach.

6.6 User interfaces

The outcome of our opinion-mining algorithm can be then denoted as: $\{<feature_i, opinion_1, opinion_2, \ldots, opinion_n>\}$

Table 5 The results of testing different combinations of feature functions in FRFs model

Combination	Precision (%)	Recall (%)	*F* score (%)
$1 + 2 + 3$	87.9	76.8	81.98
$1 + 2 + 3 + 4$	89.7	76.3	82.5
$1 + 2 + 3 + 5$	87.3	79.3	83.1
$1 + 2 + 3 + 6$	87.7	77.1	82.1
$1 + 2 + 3 + 7$	89.1	78.4	83.4
$1 + 2 + 3 + 4 + 5 + 6 + 7$	90.2	81.6	85.7

where $feature_i$ is the extracted ith feature and it is associated with a set of n distinct opinion words (i.e., $opinion_1$, $opinion_2$, …, $opinion_n$). The problem is then how to display these results to users in an informative and easily understandable way (which is for our research question 5):

Question 5: how could the opinion-mining results be represented on user interfaces for supporting users to make effective decisions?

For this purpose, we have utilized "tag cloud" visualization method. In recent years, tag cloud (or called word cloud) has popularly appeared in social networking sites to depict user-generated tags or describe the word content of web sites. The importance of a tag is reflected with the font size. For example, bigger size represents that the tag was more frequently used by users to annotate an item. Given that the tag cloud has been shown advantageous in presenting descriptive information (e.g., summarizing web search results), in reducing user frustration (Kuo et al. 2007), and in assisting users in retrieving items (such as images) (Callegari and Morreale 2010), we believe that it can help present the opinion values.

Concretely, Fig. 10 shows a sample interface design in our system, which visualizes extracted features and their associated opinions in the cloud format. The interface is displayed in comparison matrix when users compare two or more products (a similar tag cloud is also attached to each product at its detail page). The opinion features are first divided into two categories: *negative* and *positive*. The reason of placing negative opinions above is motivated by our user study's implications (see Sect. 5.2). In each

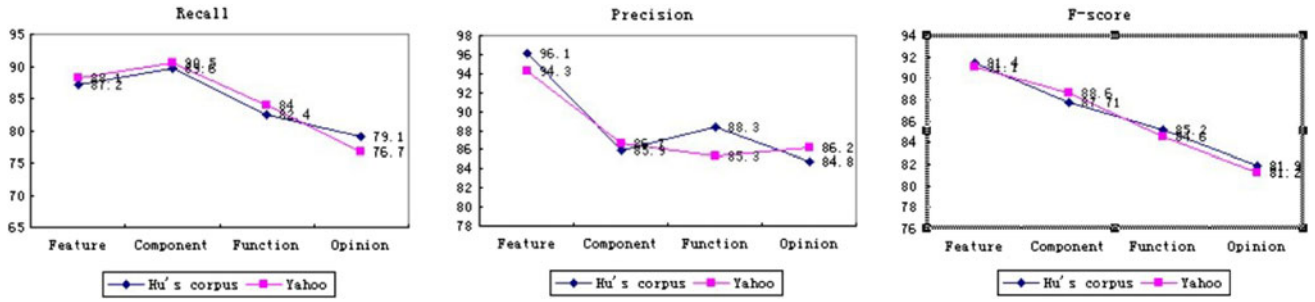

Fig. 9 Recall, precision and *F* score resulted from training CRFs with two different datasets, respectively

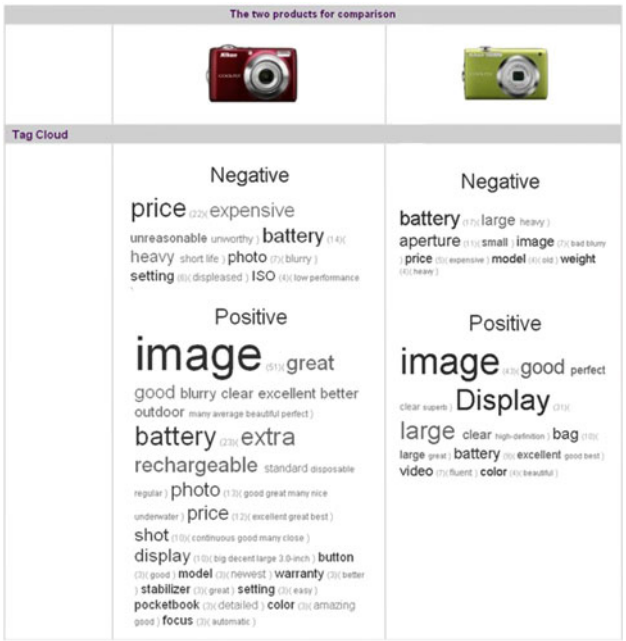

Fig. 10 Visualization of the mined opinion features in tag cloud format (i.e., the qualitative presentation)

category, every discovered feature is along with two brackets. One gives the number showing the frequency of this negative or positive feature that appears in all user reviews relate to the product. The feature's font size is thus determined by this frequency. That is, a larger font size indicates a higher frequency. Another bracket includes opinion words associated with the feature. The font size of every opinion word is also determined by its frequency of appearing in user reviews.

In our future implementation, the word's size will be also affected by users' feature preferences, which can be either explicitly stated by users or implicitly inferred from their interaction behavior. That is, if a user has placed more interest in a feature (e.g., "image"), its size can be enlarged so as to facilitate the user to examine the feature's related opinions. In addition, every word in the cloud will be made clickable, supporting users to view the original reviews.

In addition to providing the cloud format to show opinion-mining results (that we call *qualitative presentation*), we also designed an alternative, *quantitative presentation*. In this interface, every feature is quantitatively assigned two numerical opinion scores: one score reflects its number of positive opinion words (the more, the higher) and normalized in the range of [0, +5], and another score reflects its associated negative opinion words (in the range of [−5, 0]). For example, Fig. 11 shows the numerical values in form of a bar chart when a user compares two cameras.

In our prototype system, both types of interfaces are available for users to choose, because people from different backgrounds (e.g., studying majors) may have different preferences on the information presentation (Chen and Pu 2005). In our future work, we will be interested in empirically comparing them and identifying which interface would be more effective in supporting users to make product comparison.

7 Conclusions

This paper started with understanding the roles of social content in a buyer's decision-making process when they searched for inexperienced products. Based on the understanding, we have subsequently designed the system architecture, compared different social opinion-mining algorithms, and designed user interfaces for presenting the social opinions in the system. These results well answered our originally proposed five research questions (see Sect. 3.1).

Specifically, we have first conducted an exploratory user study that in-depth tracked users' objective decision behavior and interviewed their reflective thoughts. The study identified how a three-stage process was conducted (*to Question 1*, see Sect. 4.4). Furthermore, it revealed how social features can be better utilized in developing more effective decision supports, so as to assist users in seeking for needed information at different stages (*to Question 2*, see Sect. 4.5). Specifically, several system implications

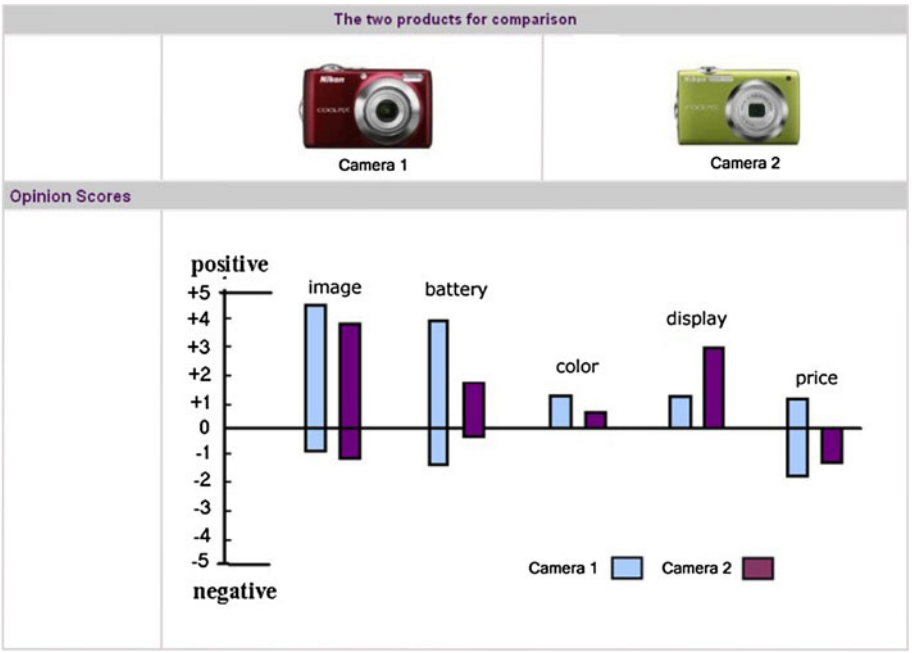

Fig. 11 Using bar chart to show numerical opinion scores (i.e., the quantitative presentation)

were concluded from the user study (*to Question 3*, see Sect. 5): (1) product popularity would better originate from usage-driven social media, when users are at the first stage in selecting interesting products. "Related Products" are also fit for this stage, and their adoption degree can be likely increased if being integrated with expert opinions and users' feature constraints; (2) when users evaluate a product in detail (Stage 2), it is suggested to provide users with extracted opinion features from user reviews, and place higher weight on negative opinions than on positive ones; (3) at the third stage of final-choice confirmation, it is also recommended to show social features' values explicitly in the comparison matrix, so as to facilitate users making more informed and effective product comparison.

Encouraged by these implications, in the follow-up work, we have in particular emphasized the processing of user-generated reviews and aimed at developing a more effective feature extraction and opinion-mining technique. Concretely, a linear-chain CRFs-based learning approach was presented in this article. Relative to related L-HMMs-based method, which assumes that each feature is independent of hidden states, CRFs-based approach can more effectively handle with interdependent features. The experiment results demonstrated the effectiveness of our opinion-mining approach in comparison with the rule-based and L-HMMs-based methods (*to Question 4*, see Sect. 6.5).

We finally presented two user interfaces that we implemented in the system in order to show the results of opinion mining (*to Question 5*, see Sect. 6.6). One was qualitative presentation based on the tag cloud format, and another is the quantitative design to indicate numerical opinion values. Both interfaces enable users to fully examine opinion features from different angles. Furthermore, meeting with system implications for Stages 2 and 3, the interfaces are not only displayed along with a product at its detail page, but also embedded into the comparison matrix to facilitate users making product comparison.

7.1 Limitations and future work

The reported work also has its limitations at several aspects. First of all, in the user study part, we selected two sites: Flickr and Yahoo Shopping, as representatives of social media sites and e-commerce sites, respectively. As we mentioned in that part, the reason of choosing them owes to their substantive and vast amount of product info. Among the various types of social and static contents that they provide, we were then able to identify users' actual information-seeking needs. However, the study did not eliminate the potential effect of confounding factors, such as site designs, on users' behavior. For this purpose, we are prepared to repeat the experiment on more sites, e.g., Amazon, Epinions, so as to further verify the results' stability.

Another limitation in our work exists in the opinion-mining algorithm evaluation. For now, we mainly compared the CRFs-based approach to the one that also belongs to model-based supervised opinion-mining branch. At a general level, it should be meaningful to identify the performance differences between model-based ones and unsupervised mining techniques, such as the work in Hu and Liu (2004). In fact, it is widely accepted that model-based learning approaches would perform better as it

involves the processes of model training and optimization (though the training effort is required). We expect in the future, through more experiments, their exact, relative merits could be clarified.

The third issue that we will address is to test our user interfaces' actual impact by means of user evaluations. In addition to comparing the two types of result presentations (i.e., qualitative presentation vs. quantitative one) as we mentioned before, we also would like to demonstrate two hypotheses. One is that embedding opinion values could effectively support users to examine a product when they decide whether the product can be a candidate or not (i.e., at their second decision stage). The second is it could further enable users to have a more accurate and confidence final choice making (i.e., at their third decision stage).

Overall, we believe that our work can be essentially contributive to resolving the demanding request: how to optimally apply user-generated content from social media environments to assist online buyers in searching inexperienced products and making decisions? Researchers from both social networking area and e-commerce decision supports could likely benefit from our user study findings and algorithm development to realize and improve their applications.

References

Adnan M, Nagi M, Kianmehr K, Tahboub R, Ridley M, Rokne J (2011) Promoting where, when and what? An analysis of web logs by integrating data mining and social network techniques to guide ecommerce business promotions. J Soc Netw Anal Min (SNAM). doi:10.1007/s13278-010-0015-3

Adomavicius G, Tuzhilin A (2005) Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans Knowl Data Eng 17(6):734–749

Al-Qaed F, Sutcliffe A (2006) Adaptive decision support system (ADSS) for B2C e-commerce. In: Proceedings of international conference on electronic commerce (ICEC'06). ACM Press, pp 492–503

Burke R, Hammond K, Young B (1997) The FindMe approach to assisted browsing. IEEE Expert Intell Syst Appl 12(4):32–40

Callegari J, Morreale P (2010) Assessment of the utility of tag clouds for faster image retrieval. In: Proceedings of the international conference on multimedia information retrieval (MIR '10), ACM, New York, pp 437–440

Chen L, Pu P (2005) Trust building in recommender agents. In: Proceedings of the workshop on web personalization, recommender systems and intelligent user interfaces at the 2nd international conference on e-business and telecommunication networks (ICETE'05), pp 135–145

Chen L, Pu P (2006) Evaluating critiquing-based recommender agents. In: Proceedings of the AAAI 2006, pp 157–162

Cialdini RB, Goldstein NJ (2004) Social influence: compliance and conformity. Annu Rev Psychol 55:591–621

Das S, Chen M (2001) Yahoo! for amazon: extracting market sentiment from stock message boards. In: Asia pacific finance association annual conference

Dave K, Lawrence S, Pennock DM (2002) Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In: Proceedings of 12th international conference on world wide web (WWW'02), pp 519–528

Engel JF, Blackwell RD, Miniard PW (1990) Consumer behavior. Dryden Press, Orlando

Fei S, Fernando P (2003) Shallow parsing with conditional random fields. In: Proceedings of the 2003 conference of the north American chapter of the association for computational linguistics on human language technology, pp 134–141

Foxall GR, Goldsmith RE, Brown S (1998) Consumer psychology for marketing. Cengage Learning Business Press

Groh G, Ehmig C (2007) Recommendations in taste related domains: collaborative filtering vs. social filtering. In: Proceedings of the 2007 international ACM conference on supporting group work (GROUP'07). ACM Press, pp 127–136

Guy I, Chen L, Zhou MX (2010) Workshop on social recommender systems. In: Proceedings of ACM international conference on intelligent user interfaces (IUI'10), pp 433–434

Hatzivassiloglou V, Wiebe JM (2000) Effects of adjective orientation and gradability on sentence subjectivity. In: Proceedings of 18th conference on computational linguistics, pp 299–305

Häubl G, Trifts V (2000) Consumer decision making in online shopping environments the effects of interactive decision aids. Market Sci 19(1):4–21

Hu M, Liu B (2004) Mining and summarizing customer reviews. In: Proceedings of 10th ACM SIGKDD international conference on knowledge discovery and data mining (SIGKDD'04), pp 168–177

Jin W, Ho H, Srihari RK (2009) OpinionMiner: a novel machine learning system for web opinion mining and extraction. In: Proceedings of ACM SIGKDD international conference on knowledge discovery and data mining (SIGKDD'09), pp 1195–1204

Kayaalp M, Özyer T, Özyer ST (2011) A mash-up application utilizing hybridized filtering techniques for recommending events at a social networking site. J Soc Netw Anal Min (SNAM) (to appear)

Kim YA, Srivastava J (2007) Impact of social influence in e-commerce decision making. In: Proceedings of international conference on electronic commerce (ICEC'07), pp 293–302

Knijnenburg BP, Willemsen MC (2009) Understanding the effect of adaptive preference elicitation methods on user satisfaction of a recommender system. In: Proceedings of ACM conference on recommender systems (RecSys'09). ACM Press, pp 381–384

Kuo BY, Hentrich T, Good BM, Wilkinson MD (2007) Tag clouds for summarizing web search results. In: Proceedings of the 16th international conference on world wide web (WWW'07). ACM, New York, pp 1203–1204

Lafferty JD, McCallum A, Pereira FC (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of international conference on machine learning, pp 282–289

Lee MK, Cheung CM, Sia CL, Lim KH (2006) How positive informational social influence affects consumers' decision of internet shopping? In: Proceedings of the 39th annual Hawaii international conference on system sciences (HICSS'06), IEEE Computer Society, vol 6

Leino J, Räihä K (2007) Case Amazon: ratings and reviews as part of recommendations. In: Proceedings of ACM conference on recommender systems (RecSys'07). ACM Press, pp 137–140

Mahmood T, Ricci F (2007) Learning and adaptivity in interactive recommender systems. In: International conference on electronic commerce (ICEC'07). ACM Press, pp 75–84

McCallum A (2003) Efficiently inducing features of conditional random fields. In: Proceedings of conference on uncertainty in artificial intelligence

McCarthy K, Reilly J, McGinty L, Smyth B (2005) Experiments in dynamic critiquing. In: Proceedings of 10th international conference on intelligent user interfaces (IUI'05), pp 175–182

Miao Q, Li Q, Zeng D (2010) Mining fine grained opinions by using probabilistic models and domain knowledge. In: Proceedings of 2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology

Mizerski R (1982) An attribution explanation of the disproportionate influence of unfavorable information. J Consumer Res 9(December):301–310

Olshavky RW, Granbois DH (1979) Consumer decision making: fact or fiction? J Consumer Res 6:93–100

Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inf Retr 2(1–2):1–135

Pang B, Lee L, Vaithyanathan S (2002) Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on empirical methods in natural language processing, pp 79–86

Payne JW, Bettman JR, Johnson EJ (1993) The adaptive decision maker. Cambridge University Press, Cambridge

Popescu A, Etzioni O (2005) Extracting product features and opinions from reviews. In: Proceedings of the conference on human language technology and empirical methods in natural language processing, human language technology conference, pp 339–346

Pu P, Chen L (2006) Integrating tradeoff support in product search tools for e-commerce sites. In: Proceedings of 6th ACM conference on electronic commerce (EC'06). ACM Press, pp 269–278

Raeder T, Chawla NV (2011) Market basket analysis with networks. J Soc Netw Anal Min (SNAM), 2011 (to appear)

Scaffidi C, Bierhoff K, Chang E, Felker M, Ng H, Jin C (2008) Red Opal: product-feature scoring from reviews. In: Proceedings of 8th ACM conference on electronic commerce (EC'08), pp 182–191

Siersdorfer S, Sizov S (2009) Social recommender systems for web 2.0 folksonomies. In: Proceedings of the 20th ACM conference on hypertext and hypermedia (Hypertext'09). ACM Press, pp 261–270

Turney PD (2002) Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of ACM SIGKDD international conference on knowledge discovery and data mining (SIGKDD'02), pp 417–424

Wilson T, Hoffmann P, Somasundaran S, Kessler J, Wiebe J, Choi Y, Cardie C, Riloff E, Patwardhan S (2005) OpinionFinder: a system for subjectivity analysis. In: Proceedings of HLT/EMNLP on interactive demonstrations, human language technology conference, pp 34–35

Yuan Q, Zhao S, Chen L, Ding S, Zhang X, Zheng W (2009) Augmenting collaborative recommender by fusing explicit social relationships. In: ACM conference on recommender systems (RecSys'09), workshop on recommender systems and the social web, New York City, NY, USA, October 22–25, 2009