Enhancement of Infrequent Purchased Product Recommendation Using Data Mining Techniques

Noraswaliza Abdullah, Yue Xu, Shlomo Geva, Mark Looi

Discipline of Computer Science
Faculty of Science and Technology
Queensland University of Technology, 2 George Street
Brisbane QLD 4000

noraswaliza.abdullah@student.qut.edu.au,
{yue.xu, shlomo.geva, m.looi}@qut.edu.au

Abstract. Recommender Systems (RS) have emerged to help users make good decisions about which products to choose from the vast range of products available on the Internet. Many of the existing recommender systems are developed for simple and frequently purchased products using a collaborative filtering (CF) approach. This approach is not applicable for recommending infrequently purchased products, as no user ratings data or previous user purchase history is available. This paper proposes a new recommender system approach that uses knowledge extracted from user online reviews for recommending infrequently purchased products. Opinion mining and rough set association rule mining are applied to extract knowledge from user online reviews. The extracted knowledge is then used to expand a user's query to retrieve the products that most likely match the user's preferences. The result of the experiment shows that the proposed approach, the Query Expansion Matching-based Search (QEMS), improves the performance of the existing Standard Matching-based Search (SMS) by recommending more products that satisfy the user's needs.

Keywords: Recommender system, opinion mining, association rule mining, user review.

1 Introduction

The large amount of information that is available on the Internet leads to an information overload problem [1]. Recommender systems (RS) have emerged to help users deal with this problem by providing product suggestions according to their needs and requirements. Nowadays, recommender systems have been widely applied by major e-commerce websites for recommending various products and serving millions of consumers [2]. However, many of the recommender systems are developed for recommending inexpensive and frequently purchased products like books, movies and music. Many of the systems that are currently available for searching infrequently purchased products like cars or houses only provide a standard

matching-based search function, whereby the system retrieves products that match exactly with the user's query. This query is normally short and does not reflect the user requirements fully. In addition, many users do not have much knowledge about the products, thus, they cannot provide detailed requirements of the product attributes or features. Therefore, a recommender system that can predict users' preferences from the initial input given by the users is needed for recommending infrequently purchased products.

Many of the current recommendation systems are developed using a collaborative filtering (CF) approach [2][3][4]. The collaborative filtering approach utilizes a large amount of ratings data or users' previous purchase data to make meaningful recommendations. This approach is not suitable for recommending infrequently purchased products because there is no previous users' purchase history or explicit ratings data about the available products, as the products are not often purchased by the users during their lifetime, and users are not able to provide ratings for products they never use. Fortunately, with the popularity of e-commerce applications for selling products on the web, users are given more opportunity to express their opinion on products they previously owned via the online merchant websites and, as a result, more and more users share reviews concerning their experience with the products. These reviews provide valuable information that can be used by recommender systems for recommending infrequently purchased products.

This paper proposes a recommender system approach that utilizes knowledge extracted from user reviews for recommending infrequently purchased products. Opinion mining and rough set association rule mining are applied to extract knowledge from the user review data to predict a user's preferences. The knowledge about user's preferences is used to expand a user's query to improve the recommendation result.

The following sections of this paper are organized as follows. First, the related work will be briefly reviewed in section 2. Then, the proposed approach will be discussed in section 3. The experimental results and evaluation will be discussed in section 4. Finally, the conclusion will be given in section 5.

2 Related Work

Recently, automatic review mining and summarization of extracting product features values from user reviews is becoming a popular research topic [5][6][7]. Review mining and summarization, also called opinion mining, aims at extracting product features on which the reviewers express their opinion and determining whether the opinions are positive or negative [7]. [5] proposed a model of feature-based opinion mining and summarization, which uses a lexicon-based method to determine whether the opinion expressed on a product feature is positive or negative. The opinion lexicon or the set of opinion words used in this method is obtained through a bootstrapping process using the WordNet. Then, [6] proposed a technique that performs better than the previous methods by using the holistic lexicon-based approach. This technique deals with context dependent opinion words and

aggregating multiple opinion words in the same sentence, which are the two main problems of the existing techniques.

Despite the growth in the number of online reviews and the valuable information that they can provide, not much work has been done on utilizing online user reviews for creating recommendations [4]. [8] employed text mining techniques to extract useful information from review comments and then mapped the review comments into the ontology's information structure, which is used by the recommender system to make recommendations. In their approach, users must input the features of the product that are most important to them and the recommendations are generated based on the features provided by the users. In contrast, our approach aims to predict users' preferences about the product features from the initial input given by them, and use the knowledge to recommend products to the users. The following section will discuss the proposed approach in detail.

3 Proposed Approach

User reviews contain written comments expressed by previous users about a particular product. Each comment contains a user's opinion or how the user feels about the product's features (e.g. good or bad). Opinion mining techniques are applied on user reviews to determine each user's sentimental orientation towards each feature, which indicates whether the user likes or dislikes the product in terms of this feature. The overall orientation of each review is also determined to summarize whether a user's opinion about the product is positive, negative or neutral. The user's opinions generated from the reviews reflect their viewpoint concerning the quality of the products. A review with a positive orientation indicates that the reviewer (i.e. the user) was satisfied with the product in some aspects. This means that at least some attributes of this product were attractive to the user. If we can identify these attractive attributes for each product, based on these attributes we can determine the products that will be of most interest to the user. Based on this idea, we propose to apply association rule mining techniques to generate patterns and association rules from users' positive reviews. By using the extracted patterns and association rules for a target user, we can predict the user's preferred product attributes and, thus, recommend products that best match the user's preferences.

The proposed recommender system approach contains three main processes: i) Opinion mining to extract a user's sentimental orientations to the product features from the user online reviews, summarizing and presenting the reviews in a structured format, ii) Rough set association rule mining to generate association rules between the product attribute values, and iii) Query expansion to expand a user's query by using association rules between product attribute values. The following sections will provide the definitions of the concepts and entities involved and the specific problems of this research. In addition, they will also explain each process in detail.

3.1 Definitions

This section first defines the important concepts and entities used in this paper and then highlights the specific problems that we aim to solve.

- **Product**

Products include any type of product or online service for which users can search for information or purchase. This paper focuses particularly on infrequently purchased products such as cars or houses. A product p can be represented by two-tuple (C,F), $C=\{c_1,c_2,...,c_n\}$ is a set of attributes representing the technical characteristics of the product defined by domain experts and $F=\{f_1,f_2,...f_m\}$ is a set of usage features representing the usage performance of the product defined by domain experts or the users of the product. The usage features are usually the aspects commented upon by the users of the product. In this paper, we assume that both the product attributes and usage features have been specified. For example, for the online car search domain on which we conducted our experiments, the following car characteristics and usage aspects were chosen as the car attributes and usage features:

C= {Make, Model, Series, Year, Engine Size, Fuel System, Fuel Consumption, Tank Capacity, Power, Torque, Body Type, Seating Capacity, Standard Transmission, Drive, Turning Circle, Kerb Weight, Dimension, Wheelbase}

F= {Comfort Practicality, Price Equipment, Under Bonnet, How Drives, Safety Security, Quality Reliability, Servicing Running Costs, Aesthetics Styling}

- **User Reviews**

For a product, there is a set of written reviews about the product given by users. Each review consists of a set of sentences comprised of a sequence of words. In many e-commerce websites, the product features to be reviewed have been specified so that users can provide their comments and opinions on each particular feature. For reviews that are not classified according to any specific feature, opinion mining techniques can be used to identify the product features that are addressed by each sentence in a review [5]. In this paper, we assume that the sentences in each review have been divided into groups, each of which consists of the sentences that talk about one feature of the product. Let $R=\{R_1,R_2,...,R_m\}$ be a review given by a user to a product, R_i is a set of sentences that are comments concerning feature f_i. By applying opinion mining techniques, which will be discussed in the next section, we can generate the user's sentimental orientation concerning each feature, denoted as $O=\{o_1,o_2,...,o_m\}$ and an overall orientation of the review O_{all}, where o_i, $O_{all}\in\{positive,negative,neutral\}$.

- **Structured review**

A structured review is a 4-tuple consisting of the sentimental orientations to a product generated from a review and the product's attributes and features, denoted as $sr=(C,F,O,O_{all})$, where C and F are the attributes and features of the product, O

and O_{all} are the sentimental orientations to the features and the overall orientation of the review, respectively. Let $SR = \{sr_1, sr_2, ..., sr_{|SR|}\}$ be a set of all structured reviews.

- **Information System**

Information system, I contains 2-tuple of information, denoted as $I = (U, A)$, where U is a set of objects, and A is a set of attributes for each object. In this paper, U is a set of structured reviews and A consists of the product attributes, features, the sentimental orientations to the features and the overall orientation of the review, i.e. $A = \{c_1, ..., c_n, f_1, ..., f_m, o_1, ..., o_m, O_{all}\}$.

The problems that we aim to solve are as follows:

i) Given a user review R on a product p, the review has to be summarized and represented in a structured review sr_i. Then from a set of all structured reviews SR, an information system I has to be generated using only reviews sr_i that have a positive or neutral overall orientation $O_{all} \in \{positive, neutral\}$.

ii) From the information model I, a set of association rules between product attribute values c_i has to be extracted using rough set association rule mining to represent users' preferences.

iii) To develop a query expansion technique by utilizing association rules extracted from information model I, to retrieve products that best meet the users' preferences.

3.2 Opinion Mining

We adopted the approach proposed by [5] to perform the opinion mining process. The first task in this process is to identify the sentimental orientations concerning the features. A user expresses a positive, negative or neutral opinion o_i, on each feature f_i, in a review R_i using a set of opinion words $W = \{w_1, w_2, ..., w_n\}$. To find out opinion words used by the user (e.g. good, amazing, poor etc.) that express his/her opinion on a product feature f_i, all adjectives used by the user in a review R_i are extracted. The orientation of each opinion word $ow_i \in \{negative, positive\}$ is then identified by utilizing the adjectives synonym set and antonym set in WordNet [9]. In WordNet, adjectives share the same orientation as their synonym and opposite orientations as their antonyms. To predict the orientation ow_i of a target adjective word w_i, a set of common adjectives with known orientation $S = \{s_1, s_2, ..., s_n\}$ called as seed adjectives, and WordNet are searched to find the word's synonym or antonym with the known orientation. If the word's synonym is found, the word's orientation is set to the same orientation as its synonym and the seed list is updated. Otherwise, if the word's antonym is found, the word's orientation is set to the opposite of the antonym and is added to the seed list. The process is repeated for the target words

with unknown orientation and the words' orientations are identified using the updated seed list. Finally, the sentimental orientation o_i of each feature f_i is identified by finding the dominant orientation of the opinion words in the sentence through counting the number of positive opinion words $ow_i \in \{positive\}$ and the negative opinion words $ow_i \in \{negative\}$, for a review R_i. If the number of positive opinion words is more than the negative opinion words, the orientation o_i of the feature f_i is positive $o_i \in \{positive\}$, otherwise negative $o_i \in \{negative\}$. If the number of positive opinion words equals the negative opinion words, the orientation o_i of the feature f_i is neutral $o_i \in \{neutral\}$.

Finally, opinion summarization is performed to determine the overall orientation of each review R and represent the review in a structured review $sr_i = (C, F, O, O_{all})$. O_{all} is determined by calculating the number of positive features $o_i \in \{positive\}$, neutral features $o_i \in \{neutral\}$, and negative features $o_i \in \{negative\}$ for the review. If the number of positive features and neutral features is more than negative features, the overall orientation O_{all} for the review is positive $O_{all} \in \{positive\}$, otherwise negative $O_{all} \in \{negative\}$. If the number of positive features and neutral features is equal to the negative features, the overall orientation O_{all} for the review is neutral $O_{all} \in \{neutral\}$.

3.3 Rough Set Association Rule Mining

Standard online product search engines perform a match process to find products that satisfy a user's query, which usually consists of the product attributes or characteristics that the users are looking for. However, many users do not have sufficient knowledge about the product and may not know the exact product attributes. Therefore, the attributes in the query may not be the right attributes to query. Online user reviews are provided by users who have used the product and the opinions about the product reflects the users' viewpoints concerning the product based on their experience of using the product. The products that are positively reviewed must possess attractive attributes or characteristics that pleased their users. Based on this intuition, we propose to find the associations between the product attributes from the users' reviews that have a positive orientation. These associations can be used to predict users' preferences to product attributes. In this paper, we utilize the rough set association rule mining approach [10] to find hidden patterns in data and generate sets of association rules from the data. We chose the rough set association rule mining technique because it allows us to easily select the condition and decision attributes of the rule.

Rough set data analysis starts from a data set that is also called a decision table or an information system. In the table, each row represents an object, each column represents an attribute, and entries of the table are attribute values. An attribute can be a variable or an observation or a property, etc. As we have defined above, an information system is written as $I = (U, A)$, in this paper, U is a set of structured

reviews and A consists of the product attributes, features, and the sentimental orientations , i.e. $A = \{c_1,...,c_n, f_1,..., f_m, o_1,...,o_m, O_{all}\}$.

In this paper, the information system $I = (U, A)$ is created from the structured reviews with positive/neutral orientation. Let $sr \in SR$ be a structured review, $sr(a)$ be the value of attribute $a \in A$, $U = \{sr \mid sr \in SR, sr(O_{all}) \in \{positive, neutral\}\}$ is the set of objects in the table. The information system contains attribute values for a set of products that have received good comments from the reviewers.

The next step in rough set association rule mining is to partition the information system into two disjointed classes of attributes, called condition C and decision D attributes. The information system is then called a decision table $S = (U, C, D)$, where C and D are disjointed sets of conditions and decision attributes, respectively. The condition and decision attributes are selected from product attributes C and features F in A in the information system I. The attributes chosen as the condition are the product attributes or features that are usually provided by a user as the initial input in a query and the decision contains other attributes and features of the products. For example, for the online car search on which we conducted our experiments, the car make, model, price, etc are chosen as the condition. Then, association rules are generated from the decision table through determining the decision attributes values based on the condition attribute values. The association rules between the product attributes values show the relationship between the initial product attribute values given by a user with other product attributes in which the user may be interested. Thus, these association rules can be used to represent the user's preferences to retrieve products that will most likely fit the user's requirements.

3.4 Query Expansion

The query expansion process aims to improve the initial user's query in order to retrieve more products that might fit the user's requirements. A user's query Q is represented by a set of terms $Q = \{q_1, q_2,..., q_n\}$ that the user provides to the search engine. In the product search, the terms in the query are attribute values of the product that the user is looking for. The query, generally, is very short and lacks sufficient terms to present the user's actual preferences or needs. Query expansion involves adding new attribute values $E = \{e_1, e_2,..., e_n\}$ to the existing search terms $Q = \{q_1, q_2,..., q_n\}$ to generate an expanded query $EQ = \{eq \mid eq \in (E \cup Q)\}$. The attribute values $eq_i \in EQ$ are used to retrieve products to recommend to the user. All products that have attribute values that match with any attribute values of the expanded query eq_i are selected as the candidate products $CP = \{cp_1, cp_2,...., cp_n\}$. The similarity between each product $cp_i \in CP$ and the expanded query EQ is calculated by matching each attribute value of the product cp_i with the value of the same product attribute eq_i in the expanded query. The similarity value v_i for the product cp_i with the expanded query EQ is calculated as the total number of attribute values

of the product that match with the attribute values in the expanded query. Then, the products are ranked based on the similarity value. The top-N products are recommended to the user based on their ranking. The proposed system may retrieve products that exactly match the user's input, as well as other products in which the user may be interested by predicting the user's preferences from his/her initial input.

4 Experiment and Evaluation

4.1 Experiment Method

A case study was conducted for the cars domain. Data was collected from one of the car selling websites that contains reviews provided by users for cars previously owned by them. The dataset contains 5,504 reviews and 3,119 cars. Online reviews on cars from previous users were used to extract rules between attribute values. The opinion mining technique was first applied to generate structured reviews from the online reviews. Then ROSETTA [11], a rough set mining tool was used for extracting rules from the information system generated from the structured review. Four processes were involved in extracting rules i) Data pre-processing and attribute selection, which included preparing a dataset and selecting important attributes to be included in the decision tables, ii) Data categorization, which involved transforming the attribute values into categorical values to reduce the dimensionality of the data and to reduce the number of rules generated, iii) Selection of decision and condition attributes and iv) Rule induction, which generated rules representing associations between query terms (e.g. car make and model) and other product attributes (e.g. series year, price new, engine size, fuel system etc.) from the decision table. The example of the rule is shown below:

CarMake(Toyota), CarModel(Camry) -> Year(>2000_and<=2005), Price(>30000_and<=50000), EngineSize(>1.6L_and<=3.0L), Seat(4_5), BodyType(SEDAN), Drive(FWD), FuelSystem(FUEL_INJECTED), FuelConsumption(>9.0L_and<=11.5L), TankCapacity(>51L_and<=70L), StandardTransmission(4A), Power(>82Kw_and<=146Kw), Torque(>150Nm_and<=284Nm), TurningCircle(>11.20m), Wheelbase(>2710mm), KerbWeight(>1126Kg_and<=1520Kg), Dimension(>4794)

Two search techniques were developed - the Standard Matching-based Search (SMS) and the Query Expansion Matching-based Search (QEMS). The SMS technique retrieves cars that match with a user's query terms exactly. In addition, the QEMS technique retrieves cars based on the expanded query of the user's preferences predicted from the rules generated from rough set association rule mining process.

For evaluating the proposed approach, previous users' navigation history from the existing system log data was used as a testing data. A sequence of cars viewed by each user was generated from the log data. The first car in each user's sequence was chosen as the input and some of the attributes of the car, such as car make and car model, were considered as the query for that user. The other cars in the sequence were considered as the cars that the user is interested in and they were used as the testing cars to test whether the two search engines recommend these cars to the user. For

each query (i.e. the first car of a user), cars recommended by both systems for a different number of the top retrieved results (N=10, 20, 30, 40 and 50) were compared with the testing cars for that user. The recall and precision values for each user were calculated for both techniques by using the following formula:

$$recall = \frac{NM}{NT} \qquad \text{and} \qquad precision = \frac{NM}{NR}$$

Where *NM* is the number of cars retrieved that match with the testing cars, *NT* is the number of testing cars, and *NR* is the number of retrieved cars. Finally, the average recall and precision values for all users were calculated for both techniques.

4.2 Results

The graphs in Figure 1 and Figure 2 show the evaluation results of the proposed approach. The evaluation results show that the Query Expansion Matching-based Search (QEMS) outperformed the Standard Matching-based Search (SMS), in that this approach can retrieve more car models that the users are interested in without requiring much effort from them. The expanded query can improve the retrieval performance of the Standard Matching-based Search as it provides more keywords to represent a user's requirements or preferences.

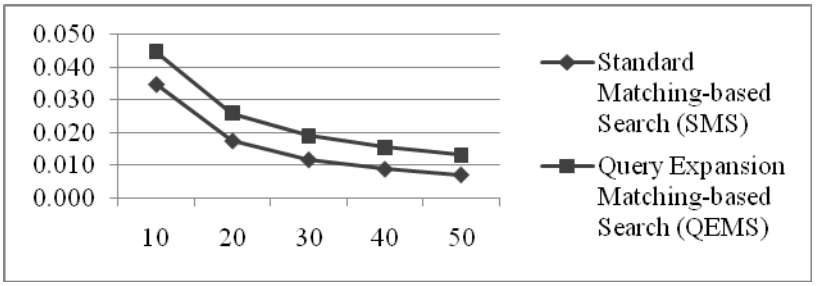

Fig. 1. Precision for different number of top retrieved results of the SMS and QEMS

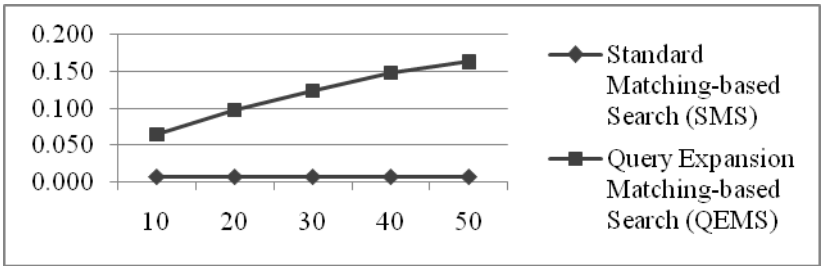

Fig. 2. Recall for different number of top retrieved results of the SMS and QEMS

5 Conclusion

We have proposed a recommender system approach for recommending infrequently purchased products by utilizing user reviews data. The evaluation result shows that our recommendation approach leads to recommendations novelty or serendipity, where more unexpected or different items that meet the users' interests will be recommended to the users. This approach is able to predict a user's preferences and may suggest more products that fit the user's requirements and, also, may help online vendors promote their products. In future work, we intend to utilize sentimental orientations of the features for improving the product recommendations.

References

1. Schafer, J. B., Konstan, J., Riedl, J.: E-commerce Recommendation Applications. Data Mining and Knowledge Discovery. 5(1-2), 115--153 (2001)
2. Leavitt, N.: Recommendation Technology: Will It Boost E-Commerce?. Computer Society. 39(5), 13--16 (2006)
3. Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Analysis of Recommendation Algorithms for E-commerce. In: 2nd ACM conference on Electronic commerce, pp. 158--167. ACM, New York (2000)
4. Aciar, S., Zhang, D., Simoff, S., Debenham, J.: Recommender System Based on Consumer Product Reviews. In: 2006 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 719--723. IEEE Computer Society, Washington (2006)
5. Hu, M., Liu, B.: Mining and Summarizing User Reviews. In: Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 168--177. ACM, New York (2004)
6. Ding, X., Liu, B., Yu, P. S.: A Holistic Lexicon-based Approach to Opinion Mining. In: International Conference on Web Search and Web Data Mining, pp. 231--240. ACM, New York (2008)
7. Zhuang, L., Jing, F., Zhu, X. Y.: Movie Review Mining and Summarization. In: 15th ACM International Conference on Information and Knowledge Management, pp. 43-50. ACM, New York (2006)
8. Aciar, S., Zhang, D., Simoff, S., & Debenham, J.: Informed Recommender: Basing Recommendations on Consumer Product Reviews. Intelligent Systems, IEEE. 22(3), 39--47 (2007)
9. Miller,G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. : Introduction to WordNet: An Online Lexical Database. International Journal of Lexicography (Special Issue). 3(4), 235--312 (1990)
10. Pawlak, Z.: Rough Sets and Intelligent Data Analysis. Information Science. 147(1-4), 1--12 (2002)
11. Øhrn, A.: ROSETTA Technical Reference Manual. Department of Computer and Information Science, Norwegian University of Science and Technology (NTNU), Trondheim, Norway (2000)

Acknowledgments. "This paper was partially supported by the Smart Services CRC (Cooperative Research Centres, Australia)".