Association for Information Systems
AIS Electronic Library (AISeL)

WHICEB 2014 Proceedings | Wuhan International Conference on e-Business

Summer 6-1-2014

Designing Ranking System for Chinese Product Search Engine Based on Customer Reviews

Samira Chaabna
School of Management, Wuhan University of Technology, China, samira.chaabna@gmail.com

Hu Wang
School of Management, Wuhan University of Technology, China

Mohammed Lutf
Department of Electronics and Information Engineering, Huazhong University of Science and Technology, China

Designing Ranking System for Chinese Product Search Engine Based on Customer Reviews

Samira Chaabna[1], Hu Wang[1], Mohammed Lutf[2]*
[1]School of Management, Wuhan University of Technology, China
[2]Department of Electronics and Information Engineering, Huazhong University of Science and Technology, China

Abstract: With the spread of e-commerce platforms, it becomes extremely difficult for the costumer to choose the right product from a large number of products, and different sellers based only on his/her own experience, product picture and meta-data. Customer's reviews present a rich source of information that have an enormous impact on the purchasing decision of the potential consumers, but reading all of the available reviews is a hard task and time consuming. Thus, the automated mining of these reviews and extract product features in order to generate a raking system present a valuable and useful tool for consumers to make well-versed decision. In this paper, we propose a product search ranking mechanism based on costumers reviews written in Chinese language. We score each product using the features extracted from the reviews. Also, a ranking function has been developed. The proposed research evaluated using customer reviews of two famous brands of mobile phones: Apple and Samsung from taobao.com. The evaluation shows a promising result compared to the existing systems.

Keywords: product search engine, customer reviews, ranking system, product feature.

1. INTRODUCTION

With the spread of e-commerce platforms, more customers are turning towards online shopping because it is convenient, fast, and reliable [1]. Given the vast variety of products, the attractive offers from different sellers, and the massive metadata of product features provided by merchants coupled with the inability to obtain the help from the professional sales staff, it becomes extremely difficult for customers to make their purchasing decisions. In order to enhance customers' satisfaction and their shopping experiences, it has become a common practice for online merchants to enable their customers to review or to express opinions on the products that have been purchased and the associated services [2,3].

In addition to comparing product features for better purchasing decision, current customers read products reviews to identify the product that satisfy their preferences. Customer reviews of a product are generally considered more honest, unbiased and comprehensive than the descriptions provided by the seller. Furthermore, reviews written by other customers describe the usage experience and perspective of (non-expert) customers with similar needs [1], but with hundreds or thousands of reviews available for some popular products, it had made hard for them to read all the reviews and to make well-versed decision on buying the products [4]. Thus, the automated mining of product reviews and opinions to produce a recalculated product ranking score is a valuable tool which would allow potential customers to make more informed decisions instead of relying only on a few of this reviews [5]. This area of study is called opinion mining or sentiment analysis [6]. Sentiment analysis of consumer reviews is a method for identifying the ways in which sentiment is expressed in text and whether such expressions include positive or negative opinions on a certain product or service [7], [8]. One of the challenges in analyzing these reviews is that the reviews contain complicated opinions on the quality of products,

* Corresponding author. Email: samira.chaabna@gmail.com (Samira Chaabna)

quality of customer services related to the sale and seller credibility, and sometimes comments unrelated to the product itself.

Usually, customers read product reviews for two reasons, either to find a product that has the best reviews regarding the product features and associated services; or in order to look for a particular review related to a specific product feature which the customer is interested in, for example a review about the battery life of a cell phone product. Thus, the main objective of this research is to automate the ordinary search process by building a product search engine based on customer reviews.

This paper illustrates how Chinese costumer reviews can be mined using techniques such as data mining and natural language processing methods to mine opinion/product features in order to generate a ranking system in product search engine based on costumers reviews. Furthermore, directing customers to products that are better matches their interests can lead to increase the usage of customer reviews-based product search engines.

The rest of this paper is organized as follows. The related works are discussed in section 2, section 3 presents the methodology of building the product search engine, and the implementation of the proposed model is described in section 4. Finally, section 5 concludes the paper and gives some perspectives for future work.

2. LITERATURE REVIEW

There have been relatively little studies focusing on product feature based ranking using customer reviews. The most closely related works on product ranking based on customer reviews are as follows:

In [12], authors proposed a ranking mechanism that ranks the reviews according to their helpfulness and their impact on sales. Their experiment results show that subjectivity analysis can give useful clues about the helpfulness of a review and about its impact on sales.

In [1], the authors proposes a feature-based product ranking mechanism, that first identify product features within a product category and analyze their frequencies and relative usage, then identify subjective and comparative sentences and assign sentiment orientations to these sentences. A weighted and directed feature graph is then built by using statistics of all review sentences. The method employs a keyword strategy to identify feature sentences and the evaluation is carried out by a pRank algorithm using Amazon.com data.

In [10], the authors developed an aspect ranking algorithm to identify the important aspects by simultaneously considering the aspect frequency and the influence of consumers' opinions given to each aspect on their overall opinions. The proposed system aims to identify important aspects of a product from consumer reviews automatically.

In [13], the authors proposed a feature-based product ranking technique by grouping users into familiar users (friends) and unfamiliar users (strangers), and assign different weights to them based on their reliability degree, where friends on the top of the list are expected to be more reliable than the rest. Then, calculate feature-based ranking list taking users' weights into consideration.

In [5], the authors introduced a product ranking system based on consumer review, considering the reviews credibility and posting date. This model have been built to help consumers to make a conclusion about the quality of particular product by developing a filtering mechanism to remove sentences/comments unrelated to the product itself from consumer reviews, such as supported services.

In [14], the authors propose a utility-based ranking mechanism on product search engines (hotels) that incorporates multidimensional consumer preferences and social media signals. Their work aims to highlight the tight linkages between user behavior on social media and search engines, by illustrating how social media can be mined and incorporated into a demand estimation model in order to generate a new ranking system in product search engines. The data collected over three months from Travelocity.com.

In [11], author proposes a ranking system that mainly focuses on identifying the product features from

customer reviews automatically instead of identifying them manually. The proposed system aims to enhance the existing ranking systems and increase the ranking accuracy.

Our work differs from theirs in the following three aspects: first this research is targeting Chinese language .Second our work considers both product quality features and the associated services features because the consumers are not looking for only product quality but also the associated services should satisfy their requirements. Third, it aims to make the search process for a particular feature within the consumer reviews or for the best product based on these reviews done automatically.

3. METHODOLOGY

An overview of the proposed system architecture is shown in Figuer.1. The system starts by crawling product pages from the web (www.taobao.com), parsing them to extract the product ID, title, meta-data and all reviews, and then build a title-based search index. At the query time, the index is used for efficient retrieval and the ranking will be according to the relevance score of each product which calculated based on the product reviews. One of the main contributions of this work is allowing the search of products with the association of a particular product feature extracted from consumer reviews, for example searching for smart phones which only have positive reviews on the battery life. The proposed system support comma separated string where the first part of the query is the search keyword and the rest of it are the product features user is interested in.

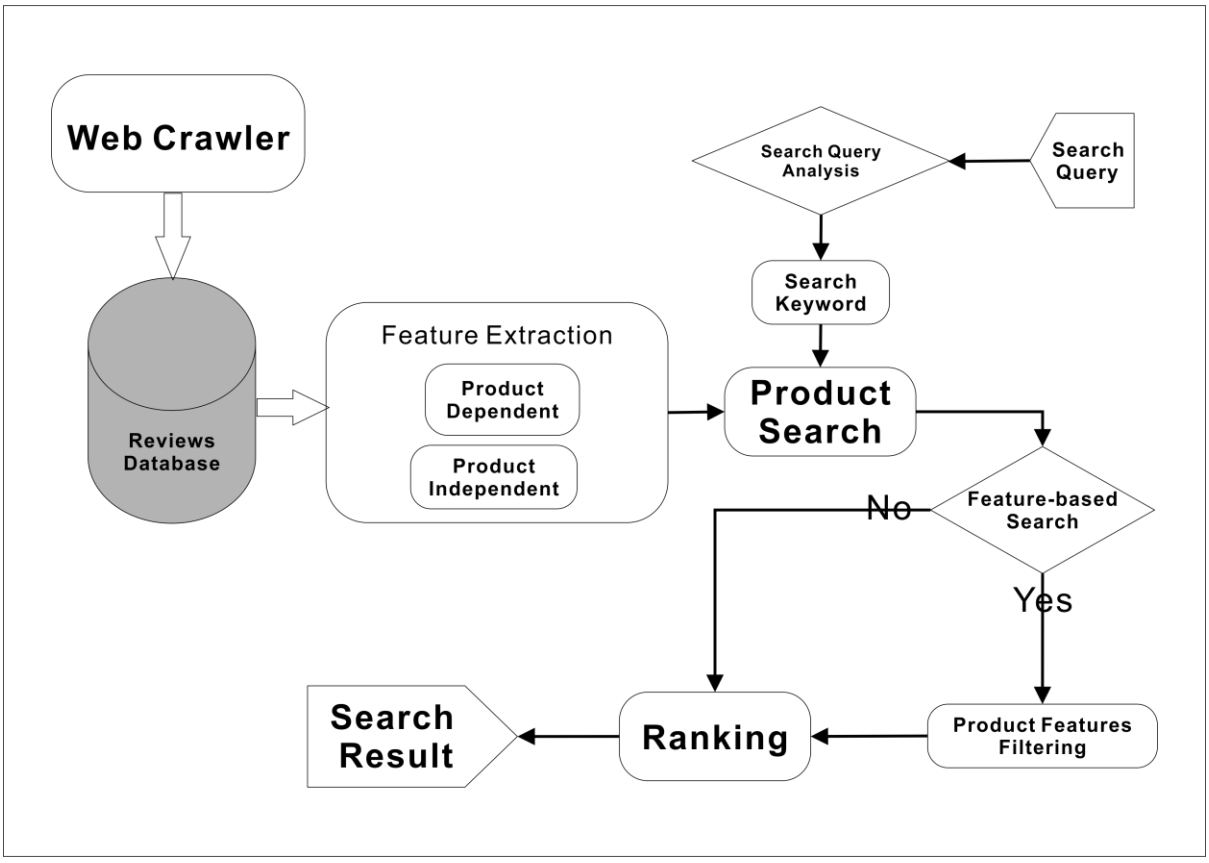

Figure 1. The proposed system architecture.

4. IMPLEMENTATION

4.1 Data collection and indexing

We collect our data from taobao.com website which is the leading e-commerce website in China. The data

collected is for smart phones of two popular brands: Apple and Samsung. About 7925 product pages have been crawled and the following information has been extracted:

- The product title: has been used to build the search index.
- Product ID: each product has a unique ID where all relevant information can be crawled based on this ID. The product ID can also be used for future updates of the product reviews.
- Product meta-data: such as the price, delivery cost, the number of deals within the last 30 days and the number of reviews.
- Product reviews: all product reviews with their supplementary information such as: posting date, the user, review usefulness.

Table 1. Summary of data collected.

Brand	Number of products	Max number of reviews	Min number of reviews
Apple	3989	12545	0
Samsung	3936	16147	0

After collecting the data, a title-based index for all crawled products has been built, because the product title always contains a full description of the product, its name, brand and model.

4.2 Features extraction

To rank the product search result, first we need to score each product. The scoring is mainly based on the product features which refer to the aspects and specifications of the product that the reviewers have expressed their opinions on. These opinions could be positive or negative. In the review text, each opinion word with one product feature form a product feature pair. Various product feature extraction techniques have been proposed by many researchers [15-18,3]. The method proposed in [15] has been seen as the elementary method for extracting product features. In our approach, the implemented method is similar to the one proposed in [18] which basically relays on extracting the typed dependencies and the collapsed dependencies for each opinion word in the review text. But, while our system was designed for Chinese language, the dependencies that describe the product features will be different from the dependencies that used with English text.

The classification of the product features may differ from one approach to

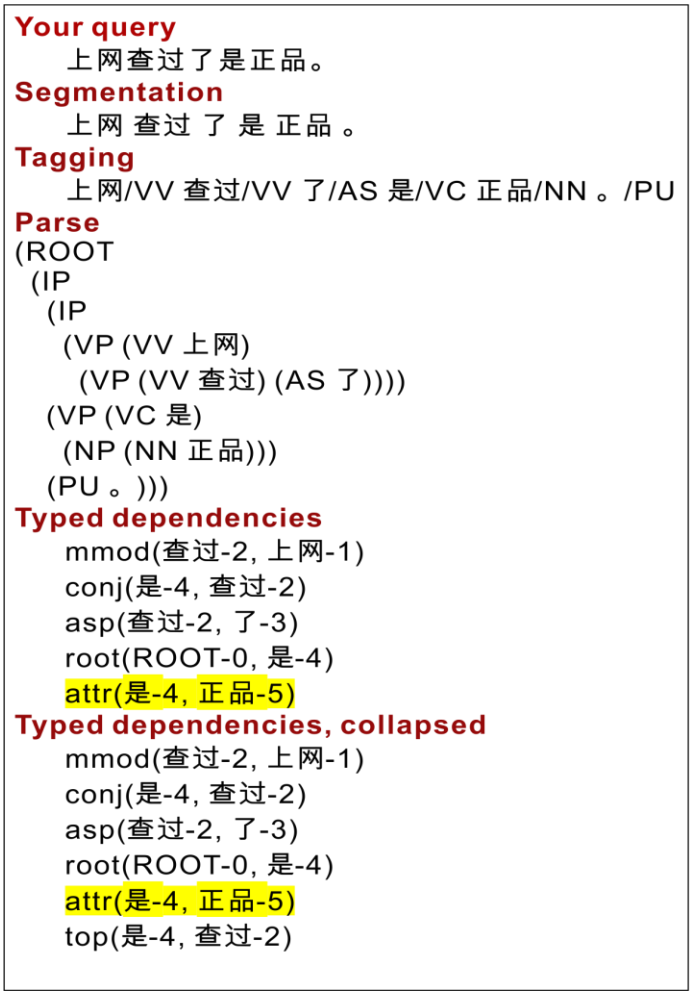

Figure 2. Example of a review for an iPhone, (The translation into English is: checked on the Internet it is original), the dependency *attr* contains the positive product features "it is original".

another. In our approach we classify the product features into two categories: product dependent features and product independent features. The product dependent features are the features that describe the product itself or its components such as the battery life of smart cell phone and the writing speed of a USB flash disk. The product independent features are the features that describe the associated services such as quality of the packaging material, speed of delivery service and after sale support. In our approach, we assign different weights for each category, assuming that the customers are more likely to be concerned about the product dependent features more than product independent features.

The first step in the processing of customer reviews is sentences segmentation. In Chinese language, there is no capitalization or space between words and which makes a Latin-based sentence segmenter produce unexpected result. Therefore, we develop a punctuation-based sentence segmenter for this task.

After segmenting each review text into individual sentences, we start parsing and tagging each word in the sentence. We used Stanford parser [19,20] (which support Chinese language) for part of speech (POS) tagging and dependencies extraction. The dependencies produced by the parser were not as accurate as we expected, but we used it as an initial estimation for extracting the product features opinion pairs. The second argument in each dependency is the opinion word, and a list of positive and negative opinion words are used as a seeds for product features pairs extraction.

For each product P_j , a product feature F_i exists if a given dependency (product feature, opinion word) is exist. The value of F_i will be equal to the difference between the positive F_{pos} and negative F_{neg} product features pairs as follow:

$$C(\mathrm{P}_j, \mathrm{F}_i) = \mathrm{count}(\mathrm{P}_j, \mathrm{F}_{i,pos}) - \mathrm{count}(\mathrm{P}_j, \mathrm{F}_{i,neg}) \tag{1}$$

And

$$C(\mathrm{P}_j, \mathrm{F}_i) = \begin{cases} C(\mathrm{P}_j, \mathrm{F}_i) & if \quad \mathrm{count}(\mathrm{P}_j, \mathrm{F}_{i,pos}) \geq \mathrm{count}(\mathrm{P}_j, \mathrm{F}_{i,neg}) \\ 0 & if \quad \mathrm{count}(\mathrm{P}_j, \mathrm{F}_{i,pos}) \prec \mathrm{count}(\mathrm{P}_j, \mathrm{F}_{i,neg}) \end{cases} \tag{2}$$

Five different dependencies are selected: nsubj (nominal subject), dobj (direct object), attr (attributive), nn (noun compound modifier), ccomp (clausal modifier). An example is shown in Figure.2 for a review of an iPhone used in our dataset (the translation of the review into English is: checked on the Internet it is original). The product feature here is it is original (是正品) and the typed dependency is attr, and since it is a positive opinion, the value $count(P_j,F_{pos})$ for the product feature pair (phone, original) will increase by 1.

4.3 Ranking

When consumers enter the search query into a product search engine, it returns a list of the matched products in a random order. To rank this list, many numerical factors could be used to score each product such as the price of the products. In our work, we propose a new scoring function which is based on the customer reviews. First we consider the products which have non-zero value for all product features pairs should have a high ranking score regardless the value of the feature product pairs. We refer to this as the product features completeness *PC*. So, for a given product *j*, if the total number of product features pairs for this product equal to *N*, then the product completeness for the product j, P_jC will be:

$$P_jC = \frac{count(\mathrm{P}_j, \mathrm{F})}{N} \tag{3}$$

where $count(P_j, F)$ = Number of non-zero product features for the product *j*. The second factor in the

scoring function is the product score P_jS which we calculate as follows:

$$P_jS = \frac{1}{M}\sum_{i=1}^{M}\frac{\mathrm{C}(\mathrm{P}_j,\mathrm{F}_i)}{\max(\mathrm{C}(\mathrm{P},\mathrm{F}_i))} \tag{4}$$

where $C(P_j, F_i) = count(P_j, F_{pos}) - count(P_j, F_{neg})$ is the value of the product feature F_i given the product j, $max(C(P,,F_i))$ = The maximum value of the product feature F_i within all products, and M is the number of features we should use for calculating the product score P_jS.

When the customer provides a search query without specifying any product feature, then we calculate the rank of the product j, P_jR according to the following function:

$$P_jR = 0.3P_jC + 0.5P_jS_{dependent} + 0.2P_jS_{independent} \tag{5}$$

If the customer provide the product feature he/she interested in, then the product score P_jR is calculated using only these product features:

$$P_jR = 0.3P_jC + 0.7P_jS_{user} \tag{6}$$

5. CONCLUSION

In this paper, we propose a novel approach to rank products by extracting different typed dependencies from the reviews of each product. The designed system is mainly built to work with reviews written in Chinese language which needs a unique processing. First, sentence segmentation is done using a punctuation-based segmenter. Then, five different dependencies are extracted from reviews to determine the product features in the reviews. The value assigned to each product feature is the difference between the count of the positive and negative product features. At the ranking step, we consider three factors in the scoring function: product features completeness score, product dependent features score and product independent features score, and we assign different weight to each factor. The effectiveness of the created system is proved through the experiment.

REFERENCES

[1] Zhang K, Narayanan R, & Choudhary A. (2010). Voice of the customers: Mining online customer reviews for product feature-based ranking. *In: Proceedings of the 3rd Conference on Online Social Networks*, pp. 11–11. USENIX Association, Berkeley, CA, USA.

[2] Hu M, & Liu B. (2004). Mining opinion features in customer reviews. *In: Proceedings of the 19th National Conference on Artifical Intelligence*, pp. 755–760. AAAI Press.

[3] Dingding W, Shenghuo Z, & Tao L. (2013). Sumview: A web-based engine for summarizing product reviews and customer opinions. *Expert Systems with Applications*, 40 (1), 27 – 33.

[4] Badina K, & Toshniwal D. (2013). Features based summarization of consumers' reviews of online products. *Procedia Computer Science*, 22, 142–151.

[5] Zhang K, Cheng Y, Liao W, & Choudhary A. (2012). Mining millions of reviews: A technique to rank products based on importance of reviews. *In: Proceedings of the 13th International Conference on Electronic Commerce*, pp. 121–128. ACM, New York, NY, USA.

[6] Liu B. (2011). Web data mining: Exploring hyperlinks, contents, and usage data. Springer-Verlag New York, Inc., Secaucus, NJ, USA,.

[7] Nasukawa T, & Yi J. (2003). Sentiment analysis: Capturing favorability using natural language processing. In: Proceedings of the 2Nd International Conference on Knowledge Capture, pp. 70–77. ACM, New York, NY, USA.

[8] Kang D, & Park Y. (2014). Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach. *Expert Systems with Applications,* 41 (4), 1041 – 1050.

[9] Liu B. (2006). Web data mining: Exploring hyperlinks, contents, and usage data. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

[10] Yu J, Zha Z J, Wang M, & Chua T S. (2011). Aspect ranking: Identifying important product aspects from online consumer reviews. *In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, pp. 1496–1505. Association for Computational Linguistics.

[11] Saranya T. (2013). Mining features and ranking products from online customer reviews. *International Journal of Engineering Research & Technology (IJERT),* 2 (10), 643 – 648.

[12] Ghose A , & Yang S. (2009). An empirical analysis of search engine advertising: Sponsored search in electronic markets. *Management Science*, 55 (10), 1605-1622.

[13] Kong R, Wang Y, Xin W, Yang T, Hu J, & Chen Z. (2011). Customer reviews for individual product feature-based ranking. *In: Proceedings of the 2011 First International Conference on Instrumentation, Measurement, Computer, Communication and Control*, pp. 449–453. IEEE Computer Society.

[14] Ghose A, Ipeirotis P G, & Li B. (2012). Designing ranking systems for hotels on travel search engines by mining user-generated and crowd-sourced content. *Management Science* 31 (3), 493 – 520.

[15] Hu M, & Liu B. (2004). Mining and summarizing customer reviews. *In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 168–177. ACM, New York, NY, USA.

[16] Kobayashi N, Iida R, Inui K, & Matsumoto Y. (2005). Opinion extraction using a learning-based anaphora resolution technique. *In: Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP- 05)*, pp. 175–180.

[17] Popescu A M, & 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*, pp. 339–346. Association for Computational Linguistics, Stroudsburg, PA, USA.

[18] Ravi K V, & Raghuveer K. (2013). Dependency driven semantic approach to product features extraction and summarization using customer reviews. *In: Advances in Computing and Information Technology*, Vol. 178, pp. 225–238.

[19] Levy R, & Manning C. (2003). Is it harder to parse Chinese, or the Chinese treebank?. *In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, pp. 439–446.

[20] Chang P C, Tseng H, Jurafsky D, & Manning C D. (2009). Discriminative reordering with Chinese grammatical relations features. *In: Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation*, pp. 51–59.