Computing (2015) 97:843–867
DOI 10.1007/s00607-015-0439-8

CAPRA: a comprehensive approach to product ranking using customer reviews

Erfan Najmi · Khayyam Hashmi · Zaki Malik · Abdelmounaam Rezgui · Habib Ullah Khan

Received: 2 April 2014 / Accepted: 19 January 2015 / Published online: 24 February 2015
© Springer-Verlag Wien 2015

Abstract Online shopping generates billions of dollars in revenues, including both the physical goods and online services. Product images and associated descriptions are the two main sources of information used by the shoppers to gain knowledge about a product. However, these two pieces of information may not always present the true picture of the product. Images could be deceiving, and descriptions could be overwhelming or cryptic. Moreover, the relative rank of these products among the peers may lead to inconsistencies. Hence, a useful and widely used piece of information is "user reviews". A number of vendors like Amazon have created whole ecosystems around user reviews, thereby boosting their revenues. However, extracting the relevant and useful information out of the plethora of reviews is not straight forward, and is a very tedious job. In this paper we propose a product ranking system that facilitates the online shopping experience by analyzing the reviews for sentiments, evaluating their usefulness, extracting and weighing different product features and aspects, ranking it among similar comparable products, and finally creating a unified rank for each product. Experiment results show the usefulness of our proposed approach in providing an effective and reliable online shopping experience in comparison with similar approaches.

E. Najmi (✉) · K. Hashmi · Z. Malik
Wayne State University, Detroit, MI 48202, USA
e-mail: erfan@wayne.edu

A. Rezgui
Department of Computer Science and Engineering, New Mexico Tech, Socorro, NM 87801, USA
e-mail: rezgui@cs.nmt.edu

H. U. Khan
Department of Accounting and Information Systems, Qatar University, 2713 Doha, Qatar
e-mail: habib.khan@qu.edu.qa

Keywords Product ranking · Sentiment analysis · Review analysis · Brand ranking · Product aspect ranking

Mathematics Subject Classification 97P99

1 Introduction

Currently more than 85 % of customers prefer online shopping to in-store shopping.[1] Major reasons for this preference are the convenience of online shopping compared to in-store shopping, and the reduced cost of storing and maintaining inventories for online retailers. Every year, the value of e-commerce trade increases exponentially. Subsequently, more categories of products are opening to customers via online shopping. The increasing use of Internet as a medium of shopping, provides an opportunity for users to express their opinions regarding their experience with products. These feedbacks show themselves on different factors on the Web as sales records, product ranks and reviews. Ghose and Ipeirotis [9] argue that reviews, their assessment of the products and their quality are effective factors that impact the sale of the products. As the different rankings of products are useful for some buyers in deciding what to buy, there are many customers who need more in-depth insight about different products. The motivation of this work is to facilitate decision making for users by creating a new rank for each product using a combination of product reviews, review ranks and the products brand rank.

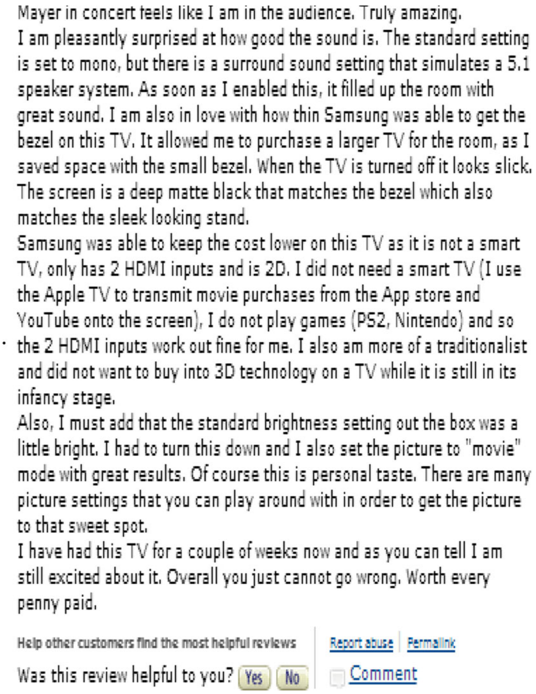
Figure 1 shows a sample review, based on user helpfulness votes, for the TV category. Using this sample and other highly voted reviews, we identify key points of online reviews (regarding their content and structure) as follows.

1. Best reviews consist of both positive and negative aspects of a product. In these kinds of reviews, it is not always possible to assign a positive or negative value to the complete review.
2. A single word in any position of a sentence can completely change the meaning and subjectivity of it, e.g., the word "but" at the beginning of the second sentence voids the negative weight of the first sentence.
3. For different features of the product the reviewers may use different terms. For example, the third paragraph of the sample review (Fig. 1) talks about picture quality, but in the first sentence the word brightness refers to the same feature using a different terminology.
4. In some cases, while the reviewer's opinion is positive in general, the review may present negative opinions at first, but later expand the discussion by providing the reasoning on why the negative points are not valid.

Our proposed approach [a comprehensive approach to product ranking (CAPRA)] starts by gathering the reviews in specific product categories. For each category, we select ten products or more with similar major features. Major features are selected after performing various analyses on reviews, and product descriptions. For example, for TVs we look at ten products with the same screen size. For each product, we

[1] http://www.safehomeproducts.com/shp2/news/news20071211.aspx.

Fig. 1 Sample review



Mayer in concert feels like I am in the audience. Truly amazing.
I am pleasantly surprised at how good the sound is. The standard setting is set to mono, but there is a surround sound setting that simulates a 5.1 speaker system. As soon as I enabled this, it filled up the room with great sound. I am also in love with how thin Samsung was able to get the bezel on this TV. It allowed me to purchase a larger TV for the room, as I saved space with the small bezel. When the TV is turned off it looks slick. The screen is a deep matte black that matches the bezel which also matches the sleek looking stand.
Samsung was able to keep the cost lower on this TV as it is not a smart TV, only has 2 HDMI inputs and is 2D. I did not need a smart TV (I use the Apple TV to transmit movie purchases from the App store and YouTube onto the screen), I do not play games (PS2, Nintendo) and so the 2 HDMI inputs work out fine for me. I also am more of a traditionalist and did not want to buy into 3D technology on a TV while it is still in its infancy stage.
Also, I must add that the standard brightness setting out the box was a little bright. I had to turn this down and I also set the picture to "movie" mode with great results. Of course this is personal taste. There are many picture settings that you can play around with in order to get the picture to that sweet spot.
I have had this TV for a couple of weeks now and as you can tell I am still excited about it. Overall you just cannot go wrong. Worth every penny paid.

Help other customers find the most helpful reviews
Was this review helpful to you? Yes No
Report abuse | Permalink
Comment

store the products' specification in the database. The specifications consist of product aspects, manufacturer, product description, and its sales rank. Sales rank is later used in comparing the results of our product rank approach with its actual sale. Next, by using the iFrame address we go through the pages of reviews and store them in the review database. Thereafter, we use *part of speech tagging* and *stemming* on the reviews. The result will be useful for sentiment analysis (SA) and review ranking. As we have different aspects for each product, different customers may have different preferences. Finding these different aspects and assigning different weights to them is what comes in the next step. In this step, we also link sentences to aspects based on the words contained in these sentences. Next we omit the unrelated pieces of information from the reviews by filtering sentences that do not correspond to the product or its aspects (features). For the next step, we assign sentiment values to these sentences consisting of negative, positive and neutral. The step-wise results are used in obtaining a final product rank (both in the general case and according to user specific preferences). Note that based on user preferences, we can prioritize the product aspects as well. Finally the user is presented with succinct, and understandable search results which assist in finding faster, personalized, and more accurate products.

Figure 2 shows the general architecture of CAPRA. Our main contributions are: (1) creating and using "Brand Rank" as a preliminary rank for new product releases. (2) We provide both aspect ranks and average product ranks, based on general user opinions and users specific needs, and finally (3) to the best of our knowledge this is

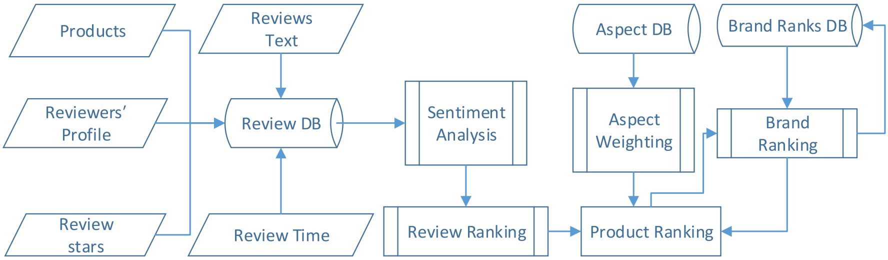


Fig. 2 Product ranking process

the first work that combines all the mentioned research fields, and creates a unified product rank.

The rest of the paper is organized as follows. In Sect. 2 we present some aspects of our data-set preparation process. Section 3 describes our approach for analysing the reviews and assigning negative, positive or neutral polarity to them. In Sect. 4 we identify different aspects of the products and introduce our approach to rank them. Section 5 outlines two different approaches to brand ranking. In Sect. 6 we describe our approach to product usefulness analysis. Section 7 summarizes all the previous sections to create a unified product rank. Experiments and results, Sect. 8, shows the result of our experiments and comparison of CAPRA to some of the similar works. Finally, Sect. 9 concludes our contributions and mentions some modifications and additions we are currently focusing on. Also in each section we first present the related work to that specific field to familiarize the reader to some of the previous work in that respective field.

2 Data-set preparation

After analyzing different resources (considering the main criteria we are looking for in the reviews; mainly descriptiveness of reviews, range of reviews from good to bad and different measures to review user experiences with products) we decided to use data-set gathered from Amazon. The positive points about the Amazon data-set are:

1. Number of reviews: on average, we have a large number of reviews for each product we considered in our data-set.
2. The Star system: the star rank each reviewer gives to the products shows the overall opinion of the reviewer of the product.
3. Review usefulness ranks: users, when deciding what to buy go through the reviews in Amazon and sometimes vote on their usefulness. These ranks are useful in defining a baseline for finding the most useful reviews.
4. Products sales rank: Amazon provides us with a sales rank number showing the sale record of products in each category.
5. Reviewer public profiles: determining the reviewer's history can assist in determining their interests, previous reviews, etc.
6. Number of replies to each review: some of the reviews have replies from other users or sometimes from the producers of a product.

While providing general users feedback on products, there are some concerns regarding Amazon reviews which we need to identify. First, for some of the reviews, portions of them do not address the product and mostly talks about the conditions for buying the product, e.g., the occasion or the time of the event. Second, the time-line of the reviews start at the release of the product and continue till the product becomes discontinued, so the user experiences are not the same over time. The third problem we have to consider is that Web sites like Amazon etc. allow different sellers to sell the same product or have different colors of the same products sold separately as a unique product. This issue not only causes duplicate products, but also provides a situation in which users repeat their reviews for different products. For the purpose of this paper, we have to remove these duplicate products and reviews. There are two general solutions for this problem. One approach runs similarity metrics on reviews using factors like bi-grams and the second approach uses reviews TF-IDF to compare their similarity. In this work, we consider reviews' bi-grams and if we find more than 80 % similarity among them, we consider these products or the reviews to be duplicates and we discard them.

Another point to mention is that we generally divide the products into two general categories, *content driven* and *use driven* products. When deciding to buy content driven products, users generally focus more on the content compared to physical attributes of the products. This approach changes users' expectations from both the products and their reviews. Hence we need to develop different processes for extracting useful information from reviews for these two types of products. The *content driven* products consist of "Books", "Music" and "Video games" etc., while the *use driven* products are the physical products for everyday use like TVs or cameras. In this paper, while the focus is on *use driven* products, in different sections we point out some difference among these two categories. For instance, analyzing *content driven* products requires modifications to each set of features we have considered in this work. We leave further work in this regard to our future work.

The process of gathering the reviews from Amazon starts by finding the products. Then, as Amazon doesn't provide users with the review texts, we scrape the Web pages to gather their text, date, reviewer info and star value of each review. The next steps focus on processing the reviews to prepare for different analysis on them. The changes consist of tokenizing, part of speech tagging and stemming. For this purpose, we use Stanford Natural Language Processing [28] Toolkit which in our assessment provides satisfactory results with acceptable efficiency [8].

The last addition to the data-set is the time token. Time of the reviews is an important factor both in comparison to the other reviews and to the release time of the products. Considering that over time newer products in the same category come out, many users review the products based on their experience with newer products. Also, in many cases, when a product has a known problem, in the newer version (of the same product), while keeping the same general specifications, companies fix the issue. While we can not provision this issue in the current work, it is an important issue to consider for future versions of our work.

Finally, because of the diversity of product categories and their reviews, we limit our product categories to HDTVs and cameras. Further discussion about our data-set and experiments is deferred until Sect. 8. We tried to gather products with review numbers in different ranges (high number of reviews, more than 100, average number, between 10 and 70, and low number of reviews, under 10) in order to consider a broader range of products.

3 Sentiment analysis

Sentiment analysis is the process of assigning polarity and sentiment values to words, sentences and the whole body of text. In recent years there has been a lot of work regarding the subject of sentiment analysis. Generally, works in this field have two main approaches. First, some works focus on assigning a positive or negative sentiment to a body of text, as a whole (examples include [22,30,33], etc.). While this approach can be useful for general text, reviews are more complicated (as shown in Sect. 1). In contrast, the second approach covers the text on a sentence-by-sentence basis [21,38]. Our premise is that, separate sections of reviews talk about positive and negative points of a product in accordance with each other, so it is logical to not consider the text as a whole and treat each sentence as a separate body of text regarding the sentiment. Moreover, for the later parts of the process, we need to consider different features of the product separately. As different portions of a review may address different aspects with different sentiments, hence we define a different approach for analysis. We divide reviews into its different aspects, and by summing up the sentiment values of each aspect, we gather the opinion of each review on the subject.

On the technical front, the SA problem is also divided into two major classes. In the first approach (lexicon based [5,30]) the focus is on creating lexicons of words, and assessing their "polarity". Polarity is defined as the orientation of the word, sentence or body of the text regarding its sentiment. Some works also store the information regarding part-of-speech taggings to be more specific about different scenarios. While this approach can be sufficient for direct and simple sentences, with the addition of complexities of natural language, it has difficulties understanding the polarity. To clarify this problem we present two examples in the following.

1. *The case of "but clause"* In most cases, the keyword "but" voids the first half of the sentence and the "but clause" can be translated alone. Similarly, there are other cases which either negate the first portion of the sentence or put more emphasis on it, e.g., "This camera size is big, but with its good design, it can easily be handled.", "I not only like the picture quality of this camera, but also its size", "This camera doesn't have a VGA port, but with internal WiFi you won't even need it". We can clearly see that identifying the differences between the sentiments of these sentences is not possible by only using the lexicon approach.
2. *The case of "negation"* In some cases, negation can make a positive polarity negative with the same weight. In other cases, however, it can change the polarity of the sentence but with less weight than the positive sentence. Moreover, in other cases negation can be used with intensifiers, which makes the behavior of the sentence unpredictable; i.e. it can decrease the weight of the polarity or completely change it with different weights. For instance, "Nobody says this is a good camera", "This camera is not very great", "In short, it is not a good camera". Similar to the previous case, differentiating between the polarities of such sentences is not possible only with the lexicon-based approach.

The main difference in lexicon based approaches is how they treat cases similar to the above mentioned examples. The general approach is to use pattern recognition to analyze these cases. While the rules and patterns introduced in these works increases

Table 1 Manual annotation result of a sample data-set

Sentence classes	% Age	Positive classes	% Age
Positive	27.2	Simple	88.2
Negative	14.08	Negated	4.2
Neutral	58.6	Complex	7.5

their accuracy to some extent, further complexities of natural language have promoted the introduction of a second approach. The text classification approach [21,33] uses classification methods to analyze and classify sentences' polarity as a whole. Normally these approaches make use of a lexicon (in some cases to be used as seed to expand and in cases as one of the classification features). Similar to these approaches to SA there are others which focus on snippets or aspect based SA [26]. The literature shows that classification approaches, specifically in more complicated texts and when implemented to specific domains has a better performance compared to lexicon based approaches.

Before going into details of our text classification approach, we expand on the complexity of reviews in the following. Our goal is to provide a more in depth analysis of reviews regarding the complexity of natural language. To understand the complexities of product reviews, we used manual annotation to classify a set of reviews in different categories of products. The aim is to find out if using a simple analyzer would suffice the needs of our sentiment analysis. Table 1 shows the results of this annotation. We separated the sentences into three main classes: neutral, positive and negative. Moreover, we divided the positive class to three sub classes. The sentences can be (1) simple positive; using simple terms to show positive opinion, e.g., “This product is amazing”. (2) Negation; to negate a negative in the sentence, e.g., “This functionality is not bad at all”. (3) Complex; which depend on the readers' knowledge to infer positive or negative meaning of a sentence, e.g., “This is like going from Blackberry to Iphone”. The results shown in Table 1 depict that around 89 % of positive sentences are simple. Thus, we can conclude that using simple analysis and negation in our classifier, we can achieve an accuracy level close to 90 %. Our analysis of online reviews reveals both structural and semantic complexities that are inherent to natural language processing.

We summarize the main issues as follows:

1. In a few cases, a negative sentence was followed by a neutral sentence. This neutral sentence was an answer to the point in the previous sentence and made it positive or vice versa. For example, “Unfortunately this product has just one HDMI port. But if you use a gaming console, that's enough”.
2. In one paragraph each sentence has neutral meaning separately but the overall theme in the paragraph has general positive or negative meaning.
3. The complication of sentences can range from a simple idioms, to comparison of two unrelated products, to an expression which does not have semantic meaning at first glance. For example “Whites are white and blacks are black” while looking completely neutral, in the TV category, this is essentially a positive attribute of the picture, and means that colors are alive and natural.

Thus, we can safely conclude that in the best case, if the system identifies all the complex sentences, clearly analyzing them would be semantically near impossible.

This is mainly due to the lack of semantic knowledge on our side to consider all the different terms of the natural language. So even if we use a more complex approach to this problem, the final accuracy would not drastically improve compared to the simple approach. Also the semantic knowledge in different categories are at least slightly different from each other which, without modification, can affect the result negatively and void the cost of the process.

In light of the above discussion, we divide the SA features into:

- Structural features: these features focus on the structure of sentences, e.g, negation.
- Semantic features: some words have gained additional and different meanings over time. This group of features focus on this concept, e.g., smileys.
- Polarity features: polarity features of words and their "pre" and "post" contexts.
- Numerical features: numbers mentioned in the sentence, e.g., 23MP.
- Review features: review features which affect the sentiment value of sentences, e.g., number of stars of the review.

Considering the polarity of words in sentences, we differentiate between "modifiers"; words which modify the polarity of a sentence, e.g., even though, and "intensifiers"; words which intensify the polarity of sentences, e.g., very. Although most of these words are grammatically adverbs and adjectives, we have to expand the list to other parts of speech (POS) as well. For example, *nothing* as a noun is generally used as a modifier. Thorough analysis and discussion of the use of modifiers and intensifiers can be found in [24]. In the field of SA there are works focusing on these words as the primary approach and follow a manual annotation of words to assign values to them [14,20]. Most of the more comprehensive approaches using this method use different compilations of the work presented in Quirk and Crystal [25]. For this work we follow a similar approach, i.e., assigning manual weights to these words which are more used in reviews compared to other parts of literature. The resulting data-set consists of 76 words following Table 2's structure. We start by annotating different phrases, from term level to whole sentence, as positive or negative. To prepare a lexicon of subjective terms we expand the SentiWordNet corpus [6], to make better sense of the phrases. SentiWordNet, itself, expands WordNet [18,19] by assigning negative and positive values to words between 0 and 1 respectively. In general, words can be *positive*, *negative*, *both* or *neutral*. An example of positive words is 'good' as in "This camera has a good picture quality". Negative words like 'negative' as in "The most negative aspect of this camera is its body size". A word which has both polarities like 'funny' as positive in "That is a very funny movie" or as negative in "The button looks funny on the TV". A neutral word is a word which does not have a specific polarity. This category consists of all the nouns or aspects of products. In our data set we make

Table 2 Shifters table sample

Word	Weight	POS
Very	2	Adj
Barely	0.5	Adv
Not	−1	Adv
Nothing	−1.5	Noun

Table 3 Neutral sentence classifier features

Word features	In subject	*Sentence features*
Words' letter case	*Modification features*	Strong/weaksubj in current sentence
Word Part-of-speech	Proceeded by adverb	Strong/weaksubj in previous sentence
Word context	Proceeded by intensifier	Strong/weaksubj in next sentence
Prior polarity	Is intensifier	Cardinal numbers in sentence
Reliability class	Modifies strongsubj	Pronoun in sentence
Review features	Modifies weaksubj	Modal in sentence
Product Category	Modified by strongsubj	Adjectives in sentence
Review star value	Modified by weaksubj	Adverbs in sentence
Structure features	Proceeded by adjective	Product aspects in sentence
In copular		Shifters in sentence
In passive		

use of the pre assigned negative or positive number of a word. The neutral words are the words which have 0 negative or positive polarity. For the other three categories we assign a threshold as show in Eq. 1 to assign positive and negative polarity to those words.

$$W_{pol} = \begin{cases} Positive & \text{if } W_p - W_n > \theta \\ Negative & \text{if } W_n - W_p > \theta \\ Neutral & \text{if } |W_p - W_n| < \theta \end{cases} \qquad (1)$$

where Wp is the positive polarity of the word, Wn is the negative polarity and θ is our assigned threshold. Wpol holds the final polarity of the word. For example, by assigning θ as .15, for the word "living" with the positivity of 0.5 and negativity of 0.125 will result in assigning positive sentiment to the word.

Our approach to SA consists of two phases. In the first phase, we solve the problem of non-neutral terms that appear in neutral sentences. As Table 1 shows we have around 59 % neutral sentences in our corpus which if not identified, because of the non-neutral terms in them, can effect the general positive and negative weights of reviews. The base classifier classifies the sentences based on the class of terms which can assign a sentiment other than neutral to unrelated sentences. To address this issue, we will use the result of the next section to remove unrelated sentences from our data-set based on the aspects in each sentence. In short, we separate the non-neutral sentences from neutral ones. The first portion, neutrality classifier, considers 27 features. These features are shown in Table 3.

The second step for the approach is polarity classification, considering that we have already removed the neutral sentences. This classifier focuses on three classes of features: Word feature, polarity features and sentence features. These features are shown in Table 4. Some of the features we used in this section have been used previously in related literature (e.g., Wilson et al. [36]). While these approaches are similar, we have tailored different parts of the approach to better suit our needs. A comparison with existing approaches is presented in Sect. 8.

Table 4 Polarity classifier features

Word features
Word's pre defined polarity: positive, negative, both, neutral
Polarity features
Negated: binary
Negated subject: binary
Modifies polarity: positive, negative, neutral, both
Modified by polarity: positive, negative, neutral, both
Conj polarity: positive, negative, neutral, both
General polarity shifter: positive, negative, very positive, very negative
Sentence features
Sentence main aspect
Emoticons in the text

4 Product aspect analyzer

Aspect analyzing (AA) is defined as extracting and analyzing products aspects and features. The subject of AA/"Topic Detection" has gained little attention in the literature, and most of the works function at the document level [29,34], as opposed to sentence level (focus of this work). NIST sponsored "Topic Detection and Tracking"[2] research track is one of the very few research tracks specifically targeted to this subject, i.e., focused on providing tools for English language speakers to access, correlate, and interpret multilingual sources of real-time information. In recent years, other than the general topic detection approaches [13], more focus has been given to specialized topic detection in specific fields, e.g. health care, etc. [17].

Topic detection at the sentence level is normally used in works which need to analyze documents at a deeper level than only the general subject of documents like review analysis. Sentence level topic detection, or "Aspect Analyzing", while harder in some aspects (limit of information in a single sentence compared to the whole body of text), is less complicated from other points of view (no need to post process and can judge each sentence independently). The main difference between sentence level and document level subject analysis is that in sentence level analysis, we have a limited set of words and sentences, and there is no given list of topics that we can map the sentences to. The former stops us from following the most common practices in this field (which is using classification [35]). Similarly, for lack of topics, we need to make a list of aspects related to different categories of the products. Moreover, each user has different priorities while looking at and/or buying a product. While these priorities can be substantially different, most customers in different categories of products are looking for specific features in their product. Thus, to analyze the reviews and break the sentences based on different categories, we need to gather the different aspects for each specific category. Therefore, instead of ranking a product as a whole, we break

[2] http://www.itl.nist.gov/iad/mig//tests/tdt/.

the product according to its different aspects. One probable solution to extract aspects is to parse the reviews to find the group of nouns and consider them as aspects of the product, based on their frequency [11]. While this approach finds all the product aspects, it also adds considerable noise in the process, which usually does not reflect the products or shoppers' opinions about them. For example, "I got this product from XYZ store, and as you know it's very expensive in there". If we follow the mentioned solution (of grouping nouns), the approach will consider XYZ as an aspect of the product, which can have a high frequency, if the store is a big distributor of the product. And based on Sect. 3, this sentence has negative polarity (the word *expensive* is negative and *very* is an intensifier which increases the negative polarity). To avoid this problem, in our work we also consider the product description as another source of aspects, as this body of text normally describes the important aspects of products focusing on the aspects that companies consider vital for their sales.

To extract aspects from the aforementioned resources, we use term frequency (TF) on groups of nouns using pattern matching. Part of these patterns are complete sentences with specific structures which in all cases have an accompanying adjective, and in some other cases we have numeric lists where each item is just a group or a single word. Note that our pattern list is not exhaustive, and by expanding the data-set (and increasing the frequency of each aspect), we can extract all the common aspects from reviews. In each category of products there are different words which directly or indirectly point to the same main aspect of a product. To consider these similar aspects as one and decrease the redundancy, for each pair of extracted aspects we measure their similarity. The solution we chose for this purpose makes use of adjectives in sentences. Our analysis shows it is common that in each category same adjectives are used to describe similar aspects. We ran a small experiment to prove this point by analyzing a small set of sentences from reviews of the same category. The results show that in 74 % of cases this theory is correct.

Following from the above mentioned point, we keep the adjectives from different aspects and compare them together. If 85 % of similar adjectives are used in comparison of two aspects, we consider the two aspects the same and store them. The result is a list of products which considers the aspect similarities. We have to note that even though our process is designed specifically to extract aspects, and compared to similar works in the literature, our process of extracting all group of nouns is better since it has less noise but we still need to improve the computational complexity of the process and the list of aspects needs to be refined. While the weighting process described later in this section will create a sorted list to be presented to the users; which in return acts a natural filter for aspects. To increase the precision of the aspect list we use expert opinions. This is specifically possible because we have limited our data-set to two categories. Nevertheless, for our future work we intend to reduce the noise, and eliminate the need for expert opinions. Since each category of products has similar aspects, we create an XML file for each category. A sample XML file for the camera category is shown in Fig. 3. In addition, there are three aspects that we consider for all the products i.e. "Delivery Time", "Packaging" and "Customer Support". The aspects files also store the synonyms for each aspect and the terms which can be used to describe these aspects. For example, for the feature "Refresh Rate" for a TV, the term in the title is described by a number followed by Hz. We store the regular expression of the term and

Fig. 3 Product aspects sample

```
<Category>
    Camera
</Category>
<Aspect>
    <Key term>
        Resolution
    </Key term>
    <Equal Terms>
        size, Picture size
    </Equal Terms>
    <Regex>
        [1-52]MP
        [1-52]Megapixel
        [1-52] Megapixel
    </Regex>
</Aspect>
```

match the pattern with numbers. Another part of the xml file stores the related terms to each aspect, e.g., for "Refresh Rate" we store terms "motion", "blur" and "picture quality". Each product in a given category should be compared to other products in the same category with similar base aspects. For example, in the process of purchasing a TV one can consider the size and the technology of the TV as its main features, so the final decision will be made based on these criteria, compared to other products with similar features. We can consider another user whose main criteria for buying a TV is its size and the price range. Our goal is to find a set of criteria for each category of products which are the most important for general users. After thorough analysis of different categories we believe the main criteria of each product can be found in the product title and its price. In this respect, it is noteworthy that the price range for each set of products can be different based on the product category. To consider this difference we create the margin automatically using the population standard deviation between product prices in a given category. For example if we have three products in one category, with respective prices of 300, 400 and 450, the margin used in this category is 62.

Finally, after extracting the aspects, using main priorities of the users, we assign different weights to different aspects. For example, picture quality in a TV is more important than its applications' execution speed. To compute the weight of each aspect we use Eq. 2. Also, based on our experiments we divided the weights of the negative results for each category from the positive ones and give them different weights respectively. For example, if for aspect one of a category we have 6 positive and 4 negative polarity in the reviews and we have 15 positive and 10 negative polarity in all the aspects of the category, the result of the equation for this example is calculated using the following equation. After calculating all the weights, we normalize the weighs so the summation of all weights in each category equals to 1.

$$\begin{aligned}&\alpha \times 6/15 + (1-\alpha) \times 4/10\\ &Aw = \alpha \times PA/P + (1-\alpha) \times NA/N \qquad (2)\end{aligned}$$

where Aw is the weight of each aspect, PA and NA are all the positive and negative repeats of the aspects, P and N are all the positive and negative sentiments of all the aspects respectively.

5 Brand ranking

Brand ranking or brand popularity analysis is an old research topic dating back to the late 1950s [23]. Most of the works to date show that brand rank is more related to brand loyalty and brand popularity than brand quality [4,32]. There are two main reasons for this. First, from a business perspective, what is important is how companies sell their products and product quality rank is not as effective on sale records as other ranks. Second, the access to information to measure brand value from a quality perspective, compared to brand popularity and loyalty, is hard to come by. While these approaches with focus on business ventures compared to end users are comprehensive, we have access to user reviews as a base of user opinions about brands and we try to create brand ranks for end users compared to businesses. We gather the brand ranks based on product quality from reviews to create a unified brand rank for each brand in each category.

Whenever a new product is released it takes a while for user reviews to start showing up. It may be possible to find some reviews in blogs, etc. but in e-commerce Web sites in general (like Amazon), as there is no user experience, there are no reviews for the product. In such a scenario, the knowledge base regarding this specific product is very small. While we don't have much information about the product, based on the history of the producer we can predict the popularity and quality of the product (which our analysis also proves). For example, in case of Samsung TVs, the star value of the products are 47 % four and half stars, 38 % four stars and around 10 % three and half stars. As this example shows most of the product reviews on same brands assign similar star values to the products. As more reviews come out, the weight of the brand rank decreases till we have enough reviews to completely nullify its effect.

Some studies have shown that brand quality can be generally measured regardless of the category of product from consumer's perspective [1]. This assumption (as shown in [1]) has two main requirements regarding the extension of the brand to different categories. One condition is the concept of "fit" category which means the previous product line of the brand has similarities to the new line. The second key point is the difficulty of extension for the company which directly relates to how similar the product categories are. These studies show that if the new category of products are too similar to the previous one, the consumers do not assign a high quality to the products. Assuming validity of these points (in all categories), and since we do not have valid similarity metrics between categories, and that we want to rank brands from a machine perspective, we separate brand ranks for different categories. For example, it is possible that a brand which produces Cameras also produces TVs, but as the quality of the products compared to other brands in the same category can be different, the ratings of these two should also be different. In the general case, the brand rank weight for content oriented products sets to zero. In the following, we discuss two approaches to calculate brand ranks. The first approach uses star and review ranks to calculate the brand rank, and the second approach makes use of PageRank [2] to calculate the brand ranks.

Since we do not have any product ranks in CAPRA in the beginning, we start by using the average stars for each category brand from Amazon. When we rank each product, the average rank of the product will consequently change the rank of the brands. Equation 3 shows CAPRA's brand rating process. The first portion of the equation creates the ratio of non-ranked products in the category and multiplies it by their star rank. This gives us an average value for the rank of the products which have not been ranked based on the reviews. The second part of the equation first finds a ratio of ranked products and multiply it to their rank to calculate the rank of the brand of already ranked products. The final score is normalized summation of these two scores, in the range of [0, 1].

$$Br_c = \frac{\frac{(TP_c - TRP_c)}{TP_c} \times avg(SV_{(}nRPCB) + TRP_c}{TP_c \times avg(RB_c)} \tag{3}$$

where Br is the brand rank of products, c is the category of the product, TP_c is the total number of products in the category c, TRP_c is the total number of ranked products in the category c, SV is the star value of the product, and nRPCB is the number of products from this specific brand which have not yet been ranked.

The second approach to the problem of ranking products makes use of the Page-rank algorithm [2] to rank brands. PageRank, in its original form, uses links between pages to approximate the value of each page. Formally, the page-rank equation in described as: "We assume page A has pages T1 $\cdots$ Tn which point to it (i.e., are citations). The parameter d is a damping factor which can be set between 0 and 1. We usually set d to 0.85. Also C(A) is defined as the number of links going out of page A."

$$PR_{(}A) = (1 - d) + d\left(\frac{PR(T_1)}{C(T_1)} + \cdots + \frac{PR(T_n)}{C(T_n)}\right) \tag{4}$$

In CAPRA, we implement the Page-Rank algorithm similarly to Eq. 4. We replace page A with brand A and pages pointing to it are replaced with reviews which mention the brand A. C(A) in our case will also be replaced with the number of brands mentioned in reviews of products from brand A. The performance of this approach is mainly related to how many of the reviewers consider mentioning other products as important for the value of their review. Based on our analysis, current reviews on Amazon do not have that many relation points to other reviews. The reviews are written mainly by the end users of the products which do not have the experience of using similar products. Hence, using this approach for brand ranking is not as useful as the previous approach. For our future work we intend to expand to reviews from the Web, and find reviews from experts in each field (which can provide information and comparison on similar products from different brands).

6 Review usefulness analysis

The 'usefulness' of a review may vary from one user to another. The concept of usefulness of reviews is effected by the reader's perspective of the product and his/her approach on product selection. Also the writing of the review, its tone, the words a writer uses, and all the other details which are not measurable using machine learned approaches may alter the usefulness. This is why even if we create a complete classifying function with 100 % accuracy from one person's perspective, there is no guarantee that someone else will have the same point of view and accepts a review as useful. Thus, we meditate on general usefulness of the reviews.

We believe that just considering the usefulness votes of users for each review is not accurate enough to be considered the only factor regarding the usefulness of the reviews as mentioned in [16]:

- Imbalance vote bias: users have a tendency to value other reviews as helpful even in cases when they are not really helpful.
- Winner circle bias: users generally vote reviews which already have positive votes as helpful compared to reviews which have not gathered as much positive feedback from other users.
- Early bird bias: early reviews of products generally get more positive votes as they have more views compared to more recent reviews.

In addition to the above mentioned problems, we have found that some reviews are not actually related to the product itself, but are based on user experiences. Such judgmental reviews usually have little or no effect on end-users, and just effect the rank of the products. For example, late delivery can be a reason for a reviewer to give one star to a product, while it does not say anything about the product itself. Researches show that users normally review a product when they are extremely happy or extremely angry with a product [12]. As a big number of reviews follows this "J shape" graph; meaning the highest number of reviews are either one or one and half stars or four and half or five stars (extremely dissatisfied or extremely satisfied), we cannot assume that the star values can completely be trusted as the review value. Thus, in the following we show how our system ranks the reviews and find a more accurate ranking.

To rank a product we first need to assign scores to the reviews. As mentioned previously, most existing works assign a positive or negative value to indicate reviews' helpfulness. We use machine learning regression to assign a score to reviews. Another approach to this issue would be machine learned ranking. Use of ranking, while at the final step would create a clear ranking of reviews, needs repetitive ranking with addition of more reviews to the data-set, which in turn increases complexity and redundancy of the approach.

Support vector regression (SVR) is a widely used regression method for analyze helpfulness of reviews. While we do not differ between helpfulness and usefulness on higher levels for users, we propose that helpfulness analysis is trying to measure how helpful reviews are from end users' perspective. On the other hand, usefulness analysis targets how useful reviews are for machine analysis and product ranking. Kim et al. [15] present an approach to analyze reviews' helpfulness using SVR. The contribution of this work is not only the use of regression but also analyzing different set of features which have the best performance in this field. Their analysis shows that the best set of features consists of unigram, length, and star values of the reviews. Considering the differences we mentioned between our works, review usefulness analysis, the aforementioned work, and review helpfulness analysis, we create a separate set of features which we believe are more related and applicable for our purpose. For the kernel, radial basis function, and other settings of SVR machine, we follow Kim et al. [15] as new settings require more thorough analysis and focus on this research topic. For the purpose of training our regression, we use a training set of assigned values gathered from manually scored reviews. We use three set of scores from three different users trained to focus on important aspects reviews targeted for machine readability.

We consider two categories of features, a set of features which shows readers' point of view on how useful reviews are, for example usefulness votes of the review. The other set of features are the ones which effect the usefulness of the reviews measurable by machine, i.e. sum of aspects. Analyzing different reviews which have the highest usefulness from different Web-sites plus the related works in this field shows that the following factors are the most decisive on review usefulness:

- Length: the number of words in objective sentences is a good measure on usefulness of the reviews. A longer review usually provides more information to the users and talk about more aspects of the products (either positive or negative).
- Reviewer average rate: each reviewer has a history of other reviews. This history can be considered as history of the reviewer's reviews usefulness based on users' votes on previous reviews.
- Sum of sentiments: this feature is the total number of sentiments that has been discussed in the review (following Eq. 5).

$$SoS = \frac{(Ss_P + Ss_N)}{SS} \tag{5}$$

  where SoS is sum of sentiments, $\mathrm{Ss_P}$ and $\mathrm{Ss_N}$ are sum of positive and negative sentiments respectively, and SS is sum of all sentences in the review.
- Star value: the star rank that a reviewer assigns to a product can show how useful the review is. More extreme ranks, specially one star, can show that the reviewer is biased towards the product.
- Sum of aspects: we take into consideration the result of our aspect analysis for each review; computed as the ratio of total number of aspects in the review to the total number of aspects for the product category.
- Time of the review with respect to the release date and current date (following Eq. 6).

$$TT = e^{\frac{T_r - T_l}{T_c - T_l}} \tag{6}$$

  where $\mathrm{T_l}$, $\mathrm{T_r}$ and $\mathrm{T_c}$ are the release time of the product, the time of the review and the current time respectively.
- Spelling mistakes: we measured the number of spelling mistakes within each review using Google spell corrector,[3] and we normalized the number by dividing it to the length of the review (in characters).
- Review replies: some reviews based on their popularity have replies. This feature stores the number of replies to the review.
- Usefulness votes: the usefulness votes based on the other users' opinions.

[3] A simple Java interface for the API available in https://code.google.com/p/google-api-spelling-java/.

- Reviewer's badges: for some reviewers, Amazon assigns badges, based-on their history or their performance as a reviewer. These badges include *#1 Reviewer*, *Top 10 Reviewer*, *Top 50 Reviewer*, *Top 500 Reviewer*, *Top 1,000 Reviewer*, *Hall Of Fame Reviewer*, *Real Name Author*, *Artist*, *Manufacturer*, *Vine Voice*, etc. There are a few more badges which are not effective on user reviews quality such as *2008 Holiday Team*. We store each of these badges as a boolean value in the data-set.
- Verified Purchase: these reviews are done by users who have bought the items from Amazon. This item uses as a factor which shows the validity of the review considering the user has really bought and used the item.

For all the features, we also run a simple standard transformation to normalize and scale them to $[-1, 1]$ values [as suggested in Hsu et al. [10] to improve the performance of the support vector machines (SVM)].

7 Product ranking

The last step of our work focuses on ranking each product among similar products in its category. This entails analyzing product reviews, breaking them down and creating a ranked list of products based-on different aspects. In this regard, some works (e.g. Zhang et al. [39]) create a manual list of product aspects which are of importance for users. Then text mining techniques run on the reviews to identify subjective and comparative sentences. With this information, a graph of product aspect rankings is created. While this work is similar to our approach it has some key differences. First, the mentioned work [39] mainly focuses on comparative sentences to compare the products and rank them. In real world data-sets the number of comparative sentences is highly limited (an average less than 1 comparative sentence per review based on our analysis) which decreases the performance of this approach immensely. In contrast, we analyze any sentence available in the reviews and specifically focus on non-neutral sentences for further analysis. Second, we add brand ranking as one of the main features for product ranking which is very effective for new products or any product which does not have as many reviews as the other products in the category. Third, unlike [39] we analyze the review usefulness to filter out reviews which are not informing or useful for users. Comparison of the approach implemented in [39] to our work is presented in Sect. 8. Other works and different approaches to this problem are proposed in [7,31,37]. We consider the score of each product as a combination of the brand score (from Sect. 5), plus the score gathered from the reviews (from Sect. 6). The weight of these two variables can differ from zero to one. This number may change, based on: (i) the number of products with the same brand (in our data-set), and (ii) the number of reviews we have for this product, and the summation of word count of them. These factors assure us that even for products with no review (especially when they are new) we have a partial rate to make the product comparable to other reviews. Equation 8 shows these two factors' effects on final product rank. We consider the average number of reviews for a product in the same group (not category). Same group means products which can be considered comparable as discussed in Sect. 4. When the number of reviews of the product is more or equal to the average number it completely voids the brand rank of the product.

$$\alpha = \begin{cases} \frac{N_r}{avg(N_c)} & \text{if } N_r < avg(N_c) \\ 1 & \text{if } N_r > avg(N_c) \end{cases} \tag{7}$$

$$Pr = (1-\alpha)B_r + \alpha R_r \tag{8}$$

where N_r is number of reviews for the product, N_c is number of reviews for products in the category, B_r is brand rank of the product, and R_r is the reviews' rank of the product. The rank from the reviews follows Eq. 9. The first portion of the equation normalizes the result for products and categories with difference in number of reviews or aspects. The remaining part of the equation sums the aspects of each review to finalize the review rank.

$$R_r = \frac{1}{m \times n} \sum_{k=1}^{n} \left(\Sigma_{i=1}^{m} (Ar_{i,k} \times Rr_k) \right). \tag{9}$$

The result of this equation, product rank (PR), provides a product score. This score then will be sorted compare to other PRs. This product list is the final result of our approach which can be shown to the end user as response to their search query.

8 Experiments and results

The data-set used for our experimental contains two main categories (TVs and Cameras). We limit our data-set to these two categories considering the number of products, their reviews, and secondly, the processing limitations. One should note that as the process of analyzing reviews, sentiment analysis, aspect extraction and review ranking can be done off-line, the work load directly related to the end user is only limited to creation of the list of ranked products. We use a total of 197 products and 56,368 reviews. Our original plan for the experiments was based on 200 products which later reduced to 197 after removing the identical products. For these 197 products, we removed more than 110 reviews as they were marked identical due to having more than 80 % similarity to other reviews. A detailed specification of the data-set is shown in Table 5. The interface of the application provides the users with search options.

The search input consists of product categories, product aspects and product price range. Product category is generally the main specifications of the products which are desirable for users, such as “45in. TV”. Price range provides user with the option of selecting minimum and maximum price which is acceptable for the final products and the last component of the input screen. As many users have different priorities for

Table 5 Data-set overview

# of products	197	# of products in camera	99
# of reviews	56,368	Min # of reviews per product	0
# of products in TV	98	Max # of review per product	1,174

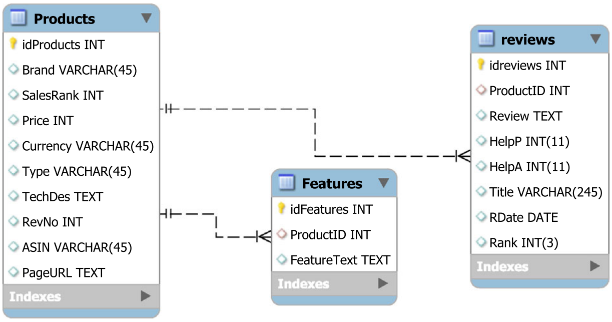


Fig. 4 Products database ER diagram

their desired aspects, product aspects search option provides users with the option to select the aspect priorities which better suits their needs. If the user does not select the important aspects, we use our default aspect weights to generate the search result. The output of the system provides five recommended products alongside the result of similar search on Amazon, omitting the products that are not in our data-set. For our experiments, we have recorded the users' choices.

The challenges for extracting reviews from Amazon arises from the limitation to the API. By the end of 2011 Amazon stopped providing API users with product reviews. Therefore to gather the reviews we had to parse the iFrame provided from Amazon and extract the reviews, their writers, number of helpful votes and their star values. As the size of the products and accordingly their reviews increases, we store the information in a database. Figure 4 shows the ER diagram of the database.

8.1 Sentiment analysis experiment result

The first experiment demonstrates the performance of the SA classifier. For both parts of our approach we used SVM with linear kernel. While content-oriented product ranking is not the focus of this paper (and hence not complete) we ran the experiment on a small set of these products as well. As the results show in Table 6, the neutrality classifier does not have the same performance on content-driven products. Apart from having a smaller training set, the increase in the number of neutral sentences (as the reviewers are more descriptive about the products), alongside using the same set of features, makes removing the neutral sentences harder. Neutral sentences effects the performance of both neutrality and polarity classifiers. As mentioned earlier our selected features are not as proficient for content-driven products as they are for the other products.

Table 6 Content oriented products sentiment analysis experiment results

	Recall	Precision	F-measure
Neutrality	45.1	53.3	48.8
Polarity	53.2	55.8	54.4

Table 7 Sentiment analysis experiment results

	TV			Camera		
	Recall	Precision	F-measure	Recall	Precision	F-measure
Neutrality	59.2	73.5	65.5	57.3	69.7	62.8
Polarity	72.5	78.4	75.3	72.3	81.4	76.5

Table 8 Neutrality and polarity classifier performance comparison

	Neutrality			Polarity		
	Recall	Precision	F-measure	Recall	Precision	F-measure
Wilson et al. [36]	48.5	58.6	54.6	64.7	67.2	65.9
CAPRA	61.6	66.9	64.1	71.6	79.5	75.3

Table 9 Polarity classifier result comparison

	Recall	Precision	F-measure
CAPRA	71.6	79.5	75.3
Sauper et al. [26]	67.6	64.2	65.8

The result in the other two categories (TVs and Cameras) are presented in Table 7, are not that different from each other. The little difference between these two categories results from the different approach of reviewers toward the products. In general, for our future work we intend to expand the categories, adding more specific features and creating an automated approach for selecting features based on the specifics of the categories.

We also provide a comparison with the Wilson et al. [36] system in the following, since it is closely related to the proposed approach. Here, we present a comparison of the result of implementing CAPRA and the mentioned work on the same data-set. As the results (Table 8) show our work and feature set has a better performance comparably. For implementation of the approach presented in Wilson et al. [36], we followed the implementation suggestion with the best performance for neutrality and polarity classifiers in [36] using respectively the TiMBL [3] tool and BoosTexter [27].

Finally we show a comparison of our classifier to another classification approach (Sauper et al. [26]). The mentioned approach works with a probabilistic topic model (mostly Dirichlet distribution) on snippets of yelp reviews, but we expand this approach to whole body of texts (to make it possible to compare the performance of the two approaches). Table 9 shows the result of this comparison. As the result shows CAPRA outperforms this approach when applied to our data-set. We believe there are two main reasons for this result. First, the mentioned approach does not consider neutral sentences in its process and second, considering the initial implementation of this system was focused on text snippets, when applied to whole reviews the system does not perform as expected.

Table 10 Result of aspect analysis in TV and camera categories

	TV	Camera
Result of pattern matching	69	46
# of aspects after similarity check	44	27
# of aspects after expert opinion	28	18

Table 11 Sample of aspect weights in the TV category

Aspects	Weights	Aspects	Weights
Picture quality	0.14	Remote backlight	0.004
Sound quality	0.08	Look	0.03
Weight	0.01	Application usability	0.04

8.2 Product aspect analyze experimental result

We ran our AA approach (defined in Sect. 4) for the mentioned category of products. In Table 10 we present our experiment result for both TV and camera category in detail. The results show the performance of aspect analyzer in extracting the aspects by providing the number of aspects for each category. The second presented information is in regards to how many aspects have been decided as synonyms and removed from the main list of aspects. Finally, we present how many of the aspects were selected using expert opinions, which is the final list presented to the end users.

Another part of aspect analyzer is assigning weights to different aspects. For the purpose of our experiment we assigned α as 0.6 to give more weight to the positive reviews of the aspects. Table 11 presents a sample of aspects weights in the TV category.

8.3 Review helpfulness and product ranking experiment result

In this section we present the result of our experiments for the final product ranking on a small number of products and reviews using both our approach and Zhang et al. [39]. Furthermore, after providing results of product ranking using steps from previous sections, we present the result of the same experiment while employing a different approach to review usefulness analysis and compare the results with CAPRA. To analyze the performance of the approach we use standard recall, precision and F-measure on the following events. For true positive class (TP) we consider when a user buys a product and is satisfied with it, false positive (FP) is when a user buys a suggested product but is dissatisfied with it, true negative (TN) is when we cannot find an appropriate product and the user agrees based on the normal returned results of the search and finally false negative is when the user is not satisfied with our recommendation, but find the desirable result from Amazon normal search. We expand the experiments with attention to 1, 3 and 5 products. Table 12 shows the result of this experiment.

While this experiment was a simulation of the real word scenario, we had a limited number of products. This limitation effected the evaluation specifically in higher

Table 12 Experiment result; 1, 3, 5 product recommendations

# of products	TP	FP	TN	FN	Precision (%)	Recall (%)	F-measure (%)
1	58	27	8	7	68.24	89.23	77.33
3	71	16	8	5	81.61	93.42	87.12
5	86	4	8	2	95.56	97.73	96.63

Table 13 Correlation result for CAPRA compared to gold standard

	Pearson correlation	Spearman correlation
TV	56.5	69.2
Camera	56.8	71.1

number of suggestions as the number of products in each category is very limited compared to the real world scenario. This effect shows itself strongly in high recall as we have been very selective on the products to be added to the database.

To compare the functionality of our work to real world data, we consider the product sales rank as the gold standard for each category. Table 13 shows the result of correlation comparison between CAPRA and the gold standard separately for TV and camera categories using Pearson and Spearman correlation metrics. While the results are satisfactory, we have to remember the goal of CAPRA is to provide better product suggestion and this product is not necessarily the best selling product in the category. Also our approach is specifically tailored for dividing each product category to sub-categories and not to rank all products in each category together.

The closest work similar to CAPRA presented here for review usefulness analysis is Kim et al. [15]. While similar, there are key differences in the selected feature-set. The main difference is, with the focus of helpfulness in the mentioned work, the gold standard is defined as Amazon helpfulness votes of users. For our work, we consider helpfulness votes as one of the features for SVR. The reason (as mentioned in Sect. 6) for this decision is that we consider the usefulness of each review for ranking products and not for informative purposes for the end-users. To present a comparison between the two sets of features we run the usefulness analysis using the features from Kim et al. and complete the product ranking using its result.

We can see that for the standard definitions of “recall”, “precision”, and “F-measure”, CAPRA shows better performance (presented in Fig. 5a–c), in retrieving relevant products to user queries specially in smaller number of returned products. In comparison, the experiment shows that Zhang et al. [39], as an example of more simplistic approach to product ranking, is not performing as well as our work specially in case of recommending less number of products, but as the number of suggested products increases, the performance difference decreases. Another effect of more suggested products, recognizing the limited number of products in the data-set, the performance of three approaches exponentially becomes closer together. In real world scenarios, with the increase in the number of products in each category, the performance would differentiate more and CAPRA would show considerably better results. Considering all the mentioned points we can safely conclude that our results are satisfactory, and

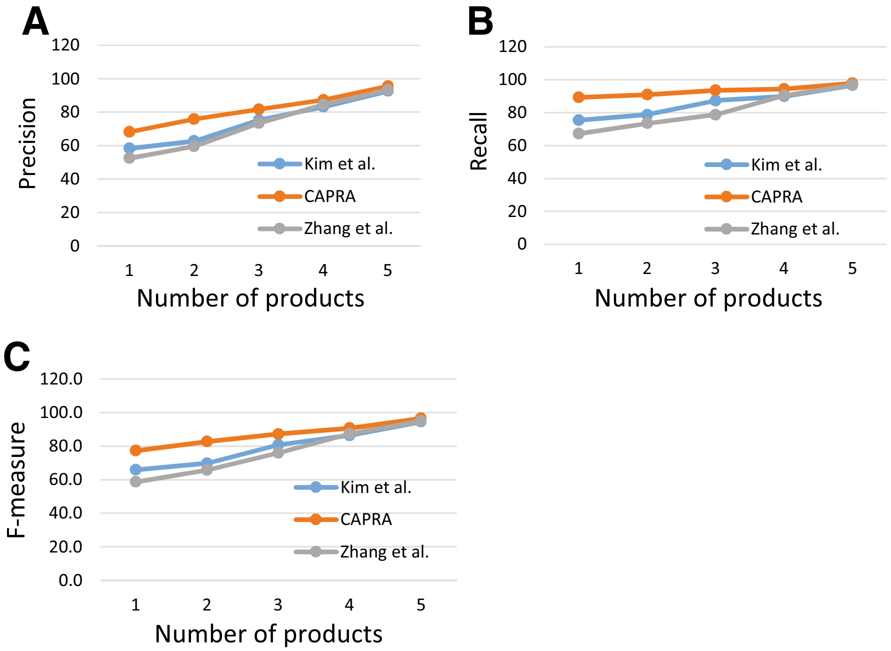


Fig. 5 CAPRA performance, compared to similar works

with small modifications and extensions (as part of the future work) can be used in real world scenarios.

9 Conclusion

In this paper we propose a novel approach to the problem of product ranking. To the best of our knowledge this is the first comprehensive approach considering different criteria of decision making for end users. The final product rank is a combination of sentiment analysis, product aspect analyzer, product brand ranks and review usefulness analysis. Moreover, the use of brand rank in the field of information retrieval for the purpose of product ranking is an innovative approach to improve the ranking process for new or less known products. Our experiments and results also are a successful proof of concept, considering the limited resources.

In our future work, we propose the following modifications and upgrades to help improve and generalize our system: We will focus on the time-line of each product and the changes, if any, made on it. This time-line history of the products effects the reviews by removing the major issues users had or the modifications the company has made on the product. The search process currently is very primeval on identifying specifications of product in user queries. One of our future steps is to make the search process easier for the users, possibly by adding natural language search option for smoother input to the system. In this paper we mostly discussed the approach implemented on use driven products. For our future work we plan on extending the work to content-driven products which will require modifying some portions of our approach. We focused on two categories of products in this work. Expansion of categories should provide a better data-set suited to analyze the product aspects, review qualities and other sections of our work. While our aspect analysis approach is comprehensive, for real world purposes with many more categories of products and their ever changing nature, decreasing the noise in identifying aspects is a viable approach to consider, and to add to the system. Amazon is a very good source for reviews. However other resources on the Web can provide a more descriptive and in certain cases technical reviews for products. Lastly, using a search engine we can limit and modify the review set using different search patterns to limit the reviews to more desirable features.

References

1. Aaker DA, Keller KL (1990) Consumer evaluations of brand extensions. J Mark 54(1):27–41
2. Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. Comput Netw ISDN Syst 30(1):107–117
3. Daelemans W, Zavrel J, Van der Sloot K, Van den Bosch A (2003) Timbl: tilburg memory-based learner. Version 4:02-01
4. Dawar N, Parker P (1994) Marketing universals: consumers' use of brand name, price, physical appearance, and retailer reputation as signals of product quality. J Mark 58(2):81–95
5. Ding X, Liu B, Yu PS (2008) A holistic lexicon-based approach to opinion mining. In: Proceedings of the international conference on Web search and web data mining. ACM, New York, pp 231–240
6. Esuli A, Sebastiani F (2006) Sentiwordnet: a publicly available lexical resource for opinion mining. Proc LREC 6:417–422
7. Feng Q, Hwang K, Dai Y (2009) Rainbow product ranking for upgrading e-commerce. Internet Comput IEEE 13(5):72–80
8. Ge SL, Song R (2010) Automated error detection of vocabulary usage in college english writing. In: IEEE/WIC/ACM international conference on Web intelligence and intelligent agent technology (WI-IAT'10), vol 3. IEEE, pp 178–181
9. Ghose A, Ipeirotis PG (2006) Designing ranking systems for consumer reviews: the impact of review subjectivity on product sales and review quality. In: Proceedings of the 16th annual workshop on information technology and systems. Citeseer, pp 303–310
10. Hsu CW, Chang CC, Lin CJ et al (2003) A practical guide to support vector classification
11. 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. ACM, New York, pp 168–177
12. Hu N, Zhang J, Pavlou PA (2009) Overcoming the J-shaped distribution of product reviews. Commun ACM 52(10):144–147
13. Joy CM, Leela S (2013) Review on sentence-level clustering with various fuzzy clustering techniques. Int J Eng 2(12):3510–3513
14. Kennedy A, Inkpen D (2006) Sentiment classification of movie reviews using contextual valence shifters. Comput Intell 22(2):110–125
15. Kim SM, Pantel P, Chklovski T, Pennacchiotti M (2006) Automatically assessing review helpfulness. In: Proceedings of the 2006 conference on empirical methods in natural language processing. Association for Computational Linguistics, pp 423–430
16. Liu J, Cao Y, Lin CY, Huang Y, Zhou M (2007) Low-quality product review detection in opinion summarization. In: Proceedings of the joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), pp 334–342
17. Lu Y, Zhang P, Liu J, Li J, Deng S (2013) Health-related hot topic detection in online communities using text clustering. PloS One 8(2):e56221
18. Miller GA (1995) Wordnet: a lexical database for english. Commun ACM 38(11):39–41
19. Miller GA, Beckwith R, Fellbaum C, Gross D, Miller KJ (1990) Introduction to wordnet: an on-line lexical database*. Int J Lexicogr 3(4):235–244

20. Nadali S, Murad MAA, Kadir RA (2010) Sentiment classification of customer reviews based on fuzzy logic. In: 2010 international symposium in information technology (ITSim), vol 2. IEEE, pp 1037–1044
21. Narayanan R, Liu B, Choudhary A (2009) Sentiment analysis of conditional sentences. In: Proceedings of the 2009 conference on empirical methods in natural language processing, vol 1. Association for Computational Linguistics, pp 180–189
22. Pang B, Lee L (2004) A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd annual meeting on association for computational linguistics. Association for Computational Linguistics, p 271
23. Pessemier EA (1959) A new way to determine buying decisions. J Mark 24(2):41–46
24. Polanyi L, Zaenen A (2006) Contextual valence shifters. In: Computing attitude and affect in text: theory and applications. Springer, New York, pp 1–10
25. Quirk R, Crystal D (1985) A comprehensive grammar of the English language, vol 6. Cambridge Univ. Press, Cambridge
26. Sauper C, Haghighi A, Barzilay R (2011) Content models with attitude. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, vol 1. Association for Computational Linguistics, pp 350–358
27. Schapire RE, Singer Y (2000) Boostexter: a boosting-based system for text categorization. Mach Learn 39(2–3):135–168
28. Stanford NLP Group (2005) Stanford parser. Retrieved 12(1):2005
29. Stoyanov V, Cardie C (2008) Topic identification for fine-grained opinion analysis. In: Proceedings of the 22nd international conference on computational linguistics, vol 1. Association for Computational Linguistics, pp 817–824
30. Taboada M, Brooke J, Tofiloski M, Voll K, Stede M (2011) Lexicon-based methods for sentiment analysis. Comput Linguist 37(2):267–307
31. Tian P, Liu Y, Liu M, Zhu S (2009) Research of product ranking technology based on opinion mining. In: Second international conference on intelligent computation technology and automation (ICICTA'09), vol 4. IEEE, pp 239–243
32. Traylor MB (1981) Product involvement and brand commitment. J Advert Res 21(6):51–56
33. Turney PD (2002) Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, pp 417–424
34. Wang X, McCallum A, Wei X (2007) Topical n-grams: phrase and topic discovery, with an application to information retrieval. In: Seventh IEEE international conference on data mining (ICDM'07). IEEE, pp 697–702
35. Wiener E, Pedersen JO, Weigend AS et al (1995) A neural network approach to topic spotting. In: Proceedings of SDAIR-95, 4th annual symposium on document analysis and information retrieval. Citeseer, pp 317–332
36. Wilson T, Wiebe J, Hoffmann P (2009) Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis. Comput Linguist 35(3):399–433
37. Zhang K, Cheng Y, Liao W, Choudhary A (2011) Mining millions of reviews: a technique to rank products based on importance of reviews. In: Proceedings of the 13th international conference on electronic commerce. ACM, New York, p 12
38. Zhang K, Cheng Y, Xie Y, Honbo D, Agrawal A, Palsetia D, Lee K, Liao W, Choudhary A (2011) SES: sentiment elicitation system for social media data. In: 2011 IEEE 11th international conference on data mining workshops (ICDMW). IEEE, pp 129–136
39. Zhang K, Narayanan R, Choudhary A (2010) Voice of the customers: mining online customer reviews for product feature-based ranking. In: 3rd workshop on online social networks