Soc. Netw. Anal. Min. (2017) 7:21
https://doi.org/10.1007/s13278-017-0442-5

ORIGINAL ARTICLE

Brand analysis framework for online marketing: ranking web pages and analyzing popularity of brands on social media

Niyati Aggrawal[1] · Archit Ahluwalia[1] · Prashi Khurana[1] · Anuja Arora[1]

Received: 11 June 2016 / Revised: 5 May 2017 / Accepted: 6 May 2017 / Published online: 18 May 2017
© Springer-Verlag Wien 2017

Abstract Online marketing is one of the best practices used to establish a brand and to increase its popularity. Advertisements are used in a better way to showcase the company's product/service and give rise to a worthy online marketing strategy. Posting an advertisement on utilitarian web pages helps to maximize brand reach and get a better feedback. Now-a-days companies are cautious of their brand image on the Internet due to the growing number of Internet users. Since there are billions of Web sites on the Internet, it becomes difficult for companies to really decide where to advertise on the Internet for brand popularity. What if, the company advertise on a page which is visited by less number of the interested (for a particular type of product) users instead of a web page which is visited by more number of the interested users?—this doubt and uncertainty—is a core issue faced by many companies. This research paper presents a Brand analysis framework and suggests some experimental practices to ensure efficiency of the proposed framework. This framework is divided into three components—(1) Web site network formation framework—a framework that forms a Web site network of a specific search query obtained from resultant web pages of three search engines-Google, Yahoo & Bing and their associated web pages; (2) content scraping framework—it crawls the content of web pages existing in the framework-formed Web site network; (3) rank assignment of networked web pages—text edge processing algorithm has been used to find out terms of interest and their occurrence associated with search query. We have further applied sentiment analysis to validate positive or negative impact of the sentences, having the search term and its associated terms (with reference to the search query) to identify impact of web page. Later, on the basis of both—text edge analysis and sentiment analysis results, we assigned a rank to networked web pages and online social network pages. In this research work, we present experiments for 'Motorola smart phone,' 'LG smart phone' and 'Samsung smart phone' as search query and sampled the Web site network of top 20 search results of all three search engines and examined up to 60 search results for each search engine. This work is useful to target the right online location for specific brand marketing. Once the brand knows the web pages/social media pages containing high brand affinity and ensures that the content of high affinity web page/social media page has a positive impact, we advertise at that respective online location. Thus, targeted brand analysis framework for online marketing not only has benefits for the advertisement agencies but also for the customers.

Keywords Brand popularity · Social network · Webpage network · Webpage ranking · Advertisement · Text edge processing · Sentiment analysis

✉ Anuja Arora
anuja.arora29@gmail.com

Archit Ahluwalia
archit.ahluwalia@gmail.com

Prashi Khurana
prashikhurana@gmail.com

[1] CSE/IT Department, Jaypee Institute of Information Technology, Noida, India

1 Introduction and related work

The rapidly increasing e-commerce sites, review analyzer sites, social networking and blogging sites have increased user-generated data volume. Beyond providing access to

these contents/reviews, these reviews/contents can be used to analyze brand popularity on these sites to post advertisements and to make advertisement most effective according to brand popularity analysis. So the important question is what does an advertisement eventually mean? Basically it is one of the means of drawing attention of the public to a product. There are many ways of advertising, be it hoardings, promotions on television, distribution of free samples, etc. The second question here would be why is an advertisement so important? Advertisements are important because they convince the customers of the product while trying to bring in new customers, enhance the image of the brand, help a new brand launch itself in public, etc. Without an advertisement, no company will be able to gain any form of publicity, which is important for any company to gain profit and do business.

Many researchers have worked on popularity analysis using comments or reviews for varying application areas (Jamali and Rangwala 2009; Siersdorfer et al. 2014; Arora et al. 2016; Bansal et al. 2016). Salman Jamali has used available comment information from digging and predicting the popularity score of linked online content using a classification and regression framework (Jamali and Rangwala 2009). Siersdorfer et al. (2014) also analyzed and mined comments and comments rating in his work and used Yahoo! News and YouTube data for this purpose. They explored the applicability of machine learning and data mining to detect acceptance of comments by the community, comments likely to trigger discussions, controversial and polarizing content, and users exhibiting offensive commenting behavior (Siersdorfer et al. 2014).

Social media comprises of online communications channels dedicated to community-based input, interaction, content sharing and collaboration. Web sites and applications dedicated to forums, micro-blogging such as twitter, social networking sites such as Facebook, Google plus, social curation sites such as Pinterest and wikis are among the different types of social media. Since 2004, the growth of social media is increasing exponentially. Today, there are about 2.3 billion active social media users which generate an expected annual growth of about 10% (https://www.brandwatch.com/2016/03/96-amazing-social-media-statistics-and-facts-for-2016/). As social media users are increasing, content posted by these users is also expanding enormously. Goeld (2014) shows the importance of online social media to collect information and other data about products, services or brands. It suggests different ways to collect, analyze and present the collected data for a variety of purposes like targeted advertising, marketing, sales and so on.

Even, we have studied various research literatures (Agarwal et al. 2011; Bansal et al. 2016) relevant to methodologies used in our work to recognize best suitable approach. Few studied works are summarized here; Agarwal et al. (2011) has analyzed sentiment of Twitter Data in his research paper and gave an approach of using sentiment analysis to classify tweets into positive, negative and neutral sentiment (Agarwal et al. 2011). He considered the prior polarity of every word using a dictionary, formed a tree kernel for every tweet by removing the stop-words and mapping each word to its part of speech tag in the dictionary and thus calculated the overall polarity of the tweet (Agarwal et al. 2011; Bansal et al. 2016). Another possible solution of sentiment analysis is Part of Speech Tagging (POS) (Agarwal et al. 2011) in which each word is represented as verb, adverb. Standard machine learning algorithms such as Naïve Bayes (6, 7), maximum entropy classification (Nigam et al. 1999) and support vector machine have also been employed for sentiment classification.

This research work proposes a framework to rank web pages (web pages are outcome of search engine for a brand-specific search query) according to their importance for a specific brand and analyzes brand popularity on social media pages using webpage ranking approach. This process initiates by generating webpage network of search query resultant web pages. Web site network idea has been taken from Merriman and O'connor (1999) and Goeldi (2014). Goeld et al. have mentioned method for targeting the delivery of advertisements over an Internet network. We have taken advantage of this (Merriman and O'connor 1999) method to form Web site network, and further webpage content has been used for advertisement purpose. We have applied text edge processing algorithm proposed by Goeldi (2014) and sentiment analysis (Spencer and Uchyigit 2012) algorithm proposed by Spencer et al. to assign page rank on the basis of webpage content similarity with the search query. Sentiment analysis was used to characterize the sentiment of webpage content as a measure to check webpage importance.

This research work aims to propose a generalized brand popularity analysis framework to measure the webpage rank and social media brand page popularity on the basis of three ranking factors.

- Assign rank to web pages on the basis of web page accessibility from three search engines and also rank linked web pages of search engine resultant web pages;
- Compute rank according to query term and its associated terms occurrence in networked web pages;
- Assign rank to search query resultant web page according to sentiments analysis of sentences which contain search query term.

The paper is organized as follows—Introduction and related work is reviewed and integrated in Sect. 1. Brand analysis framework is discussed in Sect. 2. Overall

methodology has been discussed in Sects. 3, 4 and 5. These Sects. 3, 4, and 5 discuss about Web site network formation, content scraping and rank assignment, respectively. Experimental setup and results are summarized in Sect. 6, where we present validation results of our proposed approach on real-world data including research contributions. Finally Sect. 7 concludes research work.

2 Brand analysis framework

The proposed framework has been designed to provide more relevant information for a search query with the help of text edge processing algorithm, sentiment analysis algorithm and especially by modeling network of web pages.

The proposed advertisement analysis framework has three stages as shown in Fig. 1. Figure 1 explains pictorially the same three stages as detailed below for a clear understanding of the process.

2.1 Web site network formation

Initially framework collects all the recommended pages and links from three search engines—Yahoo, Google and Bing corresponding to the search query. We have fetched the parent and the child links of all the resultant web pages. These pages sometimes have further search query term relevant web pages. The search engine resultant web pages are saved in the graph database—Neo4j to generate network of all web pages. Further, we assign page rank to all these web pages according to the mentioned page rank assignment techniques (in Sect. 2.3). For example, the search query taken to form the network is 'Motorola smart phone' (the input) while the output of this stage will be the web pages network corresponding to the search query.

2.2 Web pages content scrapping

HTML DOM parser has been used to collect data from Internet for all the networked web pages (of search query), but data thus collected are of enumerable amount, inconsistent and maximum portion is not utilizable, it is in other words, 'waste data.' Hence, it is necessary to clean the data and extract useful information only. To identify the practicable (useful) data, we have used Latent Dirichlet Allocation (LDA) python library (https://pypi.python.org/pypi/lda) and Alchemy API (http://www.alchemyapi.com/). These two serve the purpose of extracting only the required information.

2.3 Web pages rank assignment

The data which are cleaned and collected from the previous step are analyzed using algorithms like text edge processing and sentiment analysis to compute the page rank of

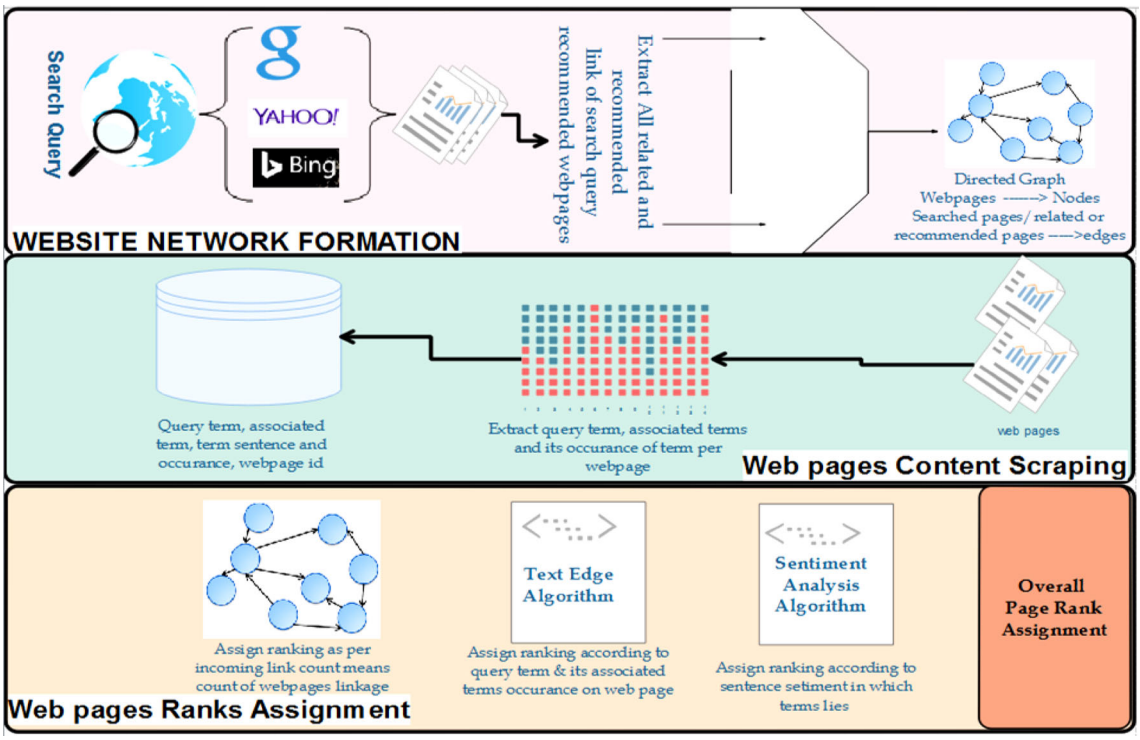

Fig. 1 Proposed brand analysis framework

entire networked web pages. The first webpage rank analysis has been done based on the number of links associated with the web pages; links which are both incoming and outgoing formed a directed graph of search query-specific Web site network.

Thereafter, text edge has been used to assign the ranking based on the query term like 'Motorola' and the other terms associated with the query-like good battery, excellent camera, etc. The sentiment analysis assigns the ranking as per sentence sentiment of the terms related to the query. The inputs for this stage are the network linked web pages, and the outputs are the assigned rank to all the web pages in the list on the basis of proposed framework.

3 Web site network formation

The initial impetus of proposed brand analysis framework is to form Web site network. This network includes entire search query resultant web pages of three search engines—Yahoo, Google & Bing. For this purpose, search query results of Google search engine are extracted using Google Search API; Yahoo and Bing search query results are crawled through HTML DOM parser.

Whenever user enters the search query, the results from the three search engines are recorded and the content on these pages stored to get more precise (to compute similarity of webpage content) list of Web sites. For a better understanding of the same, Fig. 2 gives the pictorial representation for the same (Goeldi 2014). Figure 2 is an example which suggests that the subpage3 of Yahoo has a link of subpage3 of Google in its content and thus they are linked together. Similarly subpage2 of Bing will have a link of the subpage2 of Yahoo.com. The sequence order of Subpages may vary for different search engines, which has not been considered for ranking of web pages.

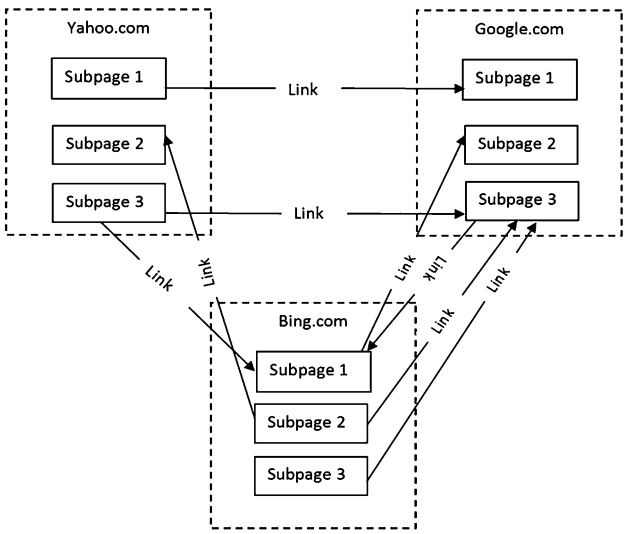

Fig. 2 Search engine crawled data extraction process

3.1 Data crawl, collection and storing in Neo4j

Yahoo and Bing search engines are crawled using HTML DOM parser library, available in PHP. The pseudocode which is used to crawl and parse data from Yahoo and Bing using HTML DOM Parser is shown in Figs. 3 and 4 respectively. Further, search query resultant web pages data are stored in the form of nodes, and each node has three different properties. These three properties are-

- Node label such as node denotes Search Engine or Web page;
- URL of linked pages;
- Page rank.

Figure 5 shows the stored nodes in Neo4j. In this figure, Search engine nodes are denoted by red color and web pages nodes by blue color. At the time of Web site network formation, page rank has been assigned on the basis of degree of node (summation of in-degree and out-degree). Further, an enhanced page rank is measured (increment/decrement) on the basis of content of that page and similarity indexing of web pages content with respect to search query and its associated terms as discussed in Sects. 4 and 5.

4 Webpage content scraping

In previous step, extracted web pages formed a network graph according to linking of one page to another as shown in Fig. 5. But it is observed that extracted pages may not have satisfactory content according to user need. Therefore, it is necessary to identify webpage content similarity with respect to search query. Henceforth, content of web pages associated with search query has been extracted using DOM Parser and stored in MongoDB database.

```
Find div tag with id=web:
Find ol tag:
Find li tag:
    Find href:
If a node already exists: Make a relation from yahoo to this node
Else:       Create a new node with the label 'page'
            Make a relation from yahoo to this node
For all : href inside this page with the required keywords
            If a node already exists
            Make a relation from the parent node to this node  else
              Create a new node with the label 'page'
            Make a relation from the parent node to this node
End for
```

Fig. 3 Yahoo search engine data crawling pseudocode

```
Find li tag with class=b_algo:
Find h2 tag:
Find a<href> tag:
If a node already exists
        Link the bing node with this node
else:   Create a new node with the label 'page'
        Make a relation from bing to this node
For all : href inside this page with the required keywords
If a node already exists
Make a relation from the parent node to this node
else
        Create a new node with the label 'page'
        Make a relation from the parent node to this node
end for
```

Fig. 4 Bing search engine data crawling pseudocode

To improve this approach further, some predefined terms (bag-of-words) are also extracted. These predefined terms are product-associated features snippets from product-specific Web sites for search query term. Consistency of predefined data is also an important measure for better accuracy. The input for this stage will be the content of networked web pages and predefined datasets. These extracted data are compared to the predefined datasets to identify the exact similarity of web page and to improve accuracy of Web site similarity for search query.

For this purpose, Python LDA API (https://pypi.python.org/pypi/lda) and Alchemy API (http://www.alchemyapi.com/) are used.

- Alchemy API generates the keyword set for the required subpage, and once the keyword set is acquired, it will be mapped to the predefined datasets. So, this alchemy API finds the query term, associated terms on the web page and maps the terms with the predefined dataset.
- Then, LDA is used to rank the web pages on the basis of term analysis which are actually useful for Web site similarity analysis.

5 Webpage rank assignment

To validate and identify similarity of web page with respect to search query or to check web page similarity to post an advertisement for a particular product on that web page, webpage rank has been assigned to these web pages. Page rank has been assigned via three ways—social graph analysis; text edge processing; sentiment analysis (Go et al. 2009).

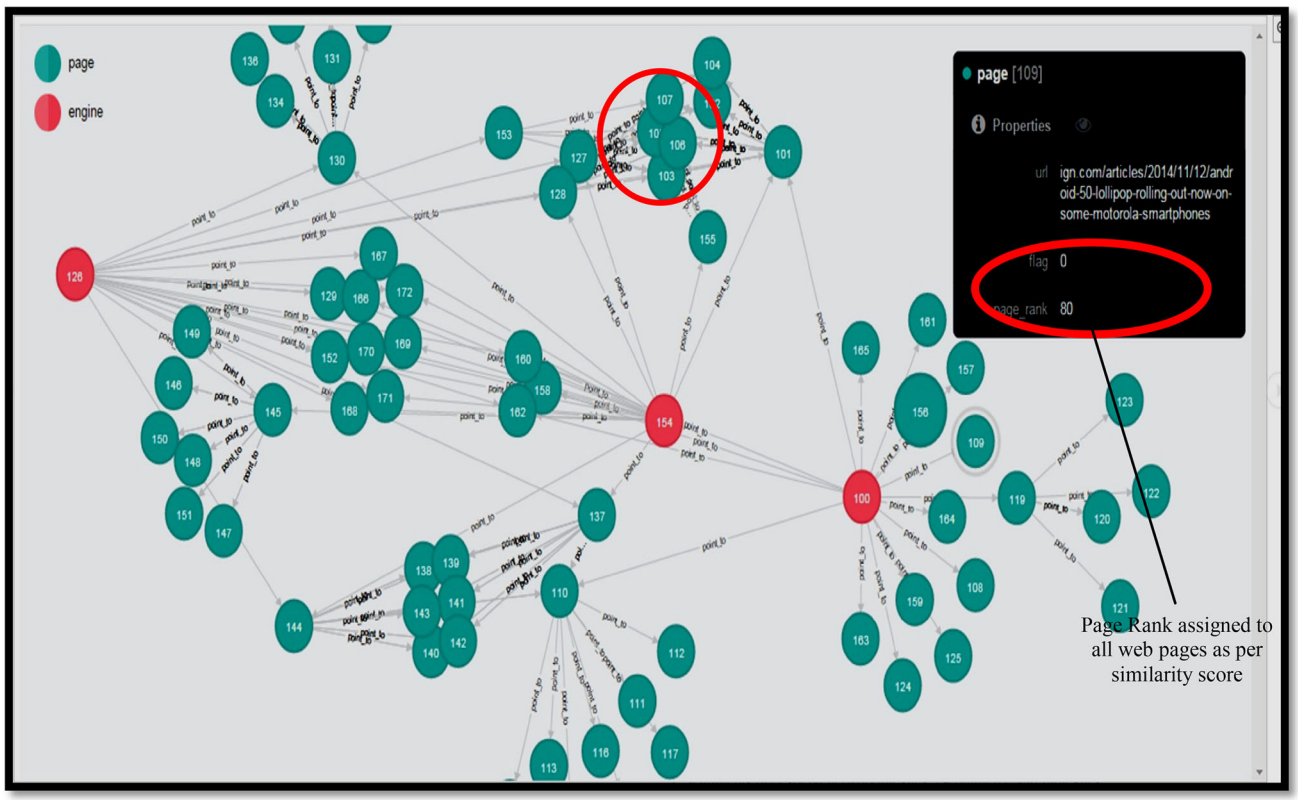

Fig. 5 Formed web pages network for 'Motorola smart phone' Search query

5.1 Social graph processing

As a graph is generated from all the webpage data, we assigned page rank to all the pages on the basis of their degree. Degree is a summation of in-degree and out-degree of web pages connectivity from the particular node (web page). Here we assigned more page rank to a node/web page if it has in-degree from search engine. A node containing highest degree is considered as most probable web pages that a user might land on.

Therefore, social graph processing-based page rank has been assigned on the basis of degree of nodes as mentioned in Eq. 1.

$$\mathrm{PR}_{\mathrm{sg}} = \sum(indegree,\ outdergree) \tag{1}$$

where $\mathrm{PR}_{\mathrm{sg}}$ is the webpage rank with respect to social graph.

5.2 Text edge processing (Goeldi 2014)

This algorithm is used to determine the frequency of occurrence of predefined terms and concepts in conjunction with the relevant brand/product. In text edge processing algorithm, terms relationship has been analyzed to determine edge relationship strength.

- Relevant sentences of web page are parsed and split into individual words using filtering and stemming technique while ignoring the stop-words and words with little information such as 'of', 'it' and 'is'. These are removed.
- Next, the relationship between the main term of interest and each found word or tuple is stored. An example to depict text edge processing is shown in Fig. 6.
- Each relationship is then counted as one instance of an "edge" between these connected objects.
- The number of relationships (example: good battery, excellent camera, etc.) between objects is added up. The resulting frequency of number of relationships is an indication of similarity of that web page.

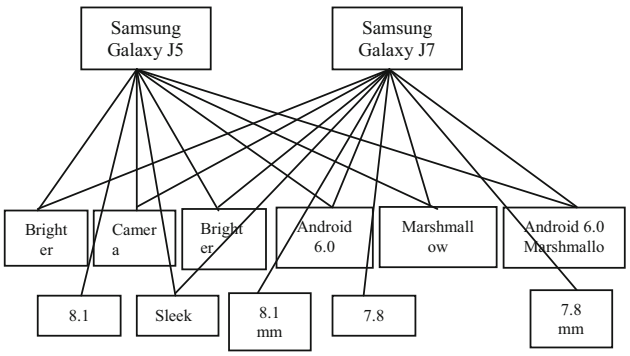

Fig. 6 Text edge processing example

Therefore, text edge processing algorithm-based page rank has been assigned on the basis of frequency of number of relationships.

$$\mathrm{PR}_{\mathrm{TE}} = \text{Frequency of number of relationships between objects of a specific node/webpage} \tag{2}$$

where $\mathrm{PR}_{\mathrm{TE}}$ is node/webpage rank with respect to text edge processing algorithm.

5.3 Sentiment analysis (Yessenov and Misailovic 2009; Pang et al. 2002)

The sentences of a web page or reviews of social networking sites tend to be longer, usually consisting of few paragraph of text, which creates problem in analyzing sentiment of sentence. For this work, considered comments are prevalently short comments like tweets, product reviews, user comments. Sentiment analysis aims to uncover the attitude of the author on a specific topic from written text. In this work, we examine the effectiveness of applying machine learning technique to the sentiment classification problem. Preprocessing techniques such as stemming, stop-word have already been applied at the time of text edge processing. Therefore, for sentiment analysis, the input dataset is preprocessed filtered data. In this work, we applied different approaches for extracting text features such as bag-of-words model, restriction of adjectives and adverbs and using WordNet synonyms knowledge. We evaluate accuracy on machine learning algorithm—Naive Bayes (Behl et al. 2014; Tan et al. 2009), which can be enhanced in future.

This algorithm is used to find whether the statement is positive negative or neutral. This algorithm has used WordStat sentiment dictionary (http://provalisresearch.com/products/content-analysis-software/wordstat-dictionary/sentiment-dictionaries/) which contains positive, negative and neutral words. We finally calculate the probability of the sentence being positive, negative or neutral based on used sentiment dictionary. Observed trends in accuracy rates are approximately 70%, but various flaws exist in used algorithm which can be resolved in future such as handling negations (words link 'not good,' etc.). Therefore, the next improvement is to calculate the probability of the sentence being negative neutral or positive at each step rather than at the end. We incorporated a technique to improve this by checking whether there exists a negation before the word. If the word next to the negation is positive, this means that the negation and the positive word have a negative impact. Similarly if we have a negation and then negative words, it will have a positive impact. We have achieved approximately 75–82% accuracy rate after considering above-mentioned negation constraints. Still few constraints are left

in negative sentences such as 'Nobody likes this product,' 'No one loves camera of this product,' which can be considered in future to improve results.

Therefore, sentiment analysis-based page rank has been assigned on the basis of positive, negative and neutral sentences about product or relevant to product for a specific node/web page.

$$\begin{aligned} PR_{SA} = {} & (\alpha * \text{ probability of number of positive sentence relevant to specific product}) \\ & + (\beta * \text{ probability of neutral sentence relevant to specific product}) \\ & - (\gamma * \text{ probability of negative sentences relevant to specific product}) \end{aligned} \quad (3)$$

where PR_{SA} is node/webpage rank with respect to sentiment analysis algorithm, and α, β and γ are weights assigned to positive, neutral and negative sentences, respectively, where $\alpha > \beta > \gamma$.

5.4 Page ranks of node/web page

A cumulative page rank has been assigned to each node/web page that exists in networked webpage graph. This rank is the summation of page rank assigned by all three used processes—social graph processing page rank (PR_{sg}), text edge processing page rank (PR_{TE}) and sentiment analysis page rank (PR_{SA}).

$$PR_c = PR_{sg} + PR_{TE} + PR_{SA} \quad (4)$$

where PR_c is cumulative page rank assigned to all web pages in the network.

6 Performance evaluation and analysis

We have taken special care in selecting the sites for evaluating performance and analyzing the brand popularity for posting advertisement.

6.1 Dataset details

To experiment the proposed advertisement analysis framework, we experimented on varying datasets. We collected dataset for varying search queries of chosen three products—Motorola, LG and Samsung from three search engines—Google, Bing and Yahoo. Also, data have been extracted for same queries from three social networking Web sites—Facebook, CNET and Flipkart.

For detailed experiment of Web site network formation, we have taken 20 search results of all three search engines, i.e., on first level 60 web pages and further, all related pages that exist on the search engine provided web pages for a upper limit 20 have been taken to compute page rank and to find out most popular pages which is most reachable/having highest affinity from all the search engines for specific product advertisement, whereas we have tested the network formation-based page rank assignment approach on 60 search results of all three search engine also and able to get satisfactory results.

Table 1 Dataset description for 'Motorola smart phone'

Social networking site—Facebook()	
~Motorola No. of posts: 185	No. of comments : 2567
~LG No. of posts: 288	No. of comments : 3561
~Samsung No. of posts: 333	No. of comments : 4651
E-commerce site—Flipkart	
~Motorola No. of pages: 22	No. of Reviews : 1212
~LG No. of pages: 51	No. of Reviews : 2076
~Samsung No. of pages: 80	No. of Reviews : 2851
Review site—CNET	
~Motorola	No. of Reviews : 987
~LG	No. of Reviews: 870
~Samsung	No. of Reviews: 1193

For brand-related discussion on social networking site, we have taken 9-week data. For this purpose, we have worked with social networking site Facebook, e-commerce site Flipkart and technical product review site CNET. Table 1 depicts the dataset details of 9-week data collected from three sites.

Some interesting analysis-based results were witnessed using proposed brand popularity analysis framework used algorithms.

6.2 Result1—highest affinity web pages list for a brand

As a result, we listed all networked web pages for search query 'Motorola Smart Phone' sorted as per the assigned cumulative page rank (PR_c). It is basically list of the web pages having highest affinity for the search query/brand. This cumulative page rank has been computed according to the proposed framework as visually depicted in Fig. 1. Proposed framework and used algorithms generated results are shown in Fig. 5 which is network of Web sites having the highest affinity for a specific brand. Table 2 presents the top 10 most probable sites according to PR_c for 'Motorola smart phone' search query.

By using this resultant list of web pages, the brand can either advertise on those web pages or ensure that the write up on that page has a positive feedback of the brand.

Table 2 Top 10 most probable web pages for search query 'Motorola smart phone'

Web page id	PR_c
gsmarena.com/motorola-phones-4.php	97
motorola-blog.blogspot.com	95
phonesreview.co.uk/category/mobile-phones/motorola/	90
motorola-blog.blogspot.com/2014/10/hello-lenovo.html	90
motorola-global-portal.custhelp.com/app/software-upgrade-news/g_id/1949	90
ign.com/articles/2014/11/12/android-50-lollipop-rolling-out-now-on-some-motorola-smartphones	80
cnet.com/topics/phones/products/motorola/	80
techradar.com/reviews/phones/mobile-phones/motorola-moto-x-1170399/review	80
techradar.com/reviews/phones/mobile-phones/moto-g-1199218/review	70
motorola.com	70

6.3 Result2: brand popularity

We have applied proposed approach—social graph analysis, text edge processing and sentiment analysis—to analyze brand popularity on chosen three web sites and compared brand popularity.

Figure 7 is a generated brand popularity comparison graph for three phone brands—Samsung, LG and Motorola. Most popular brand on all three Web sites is Samsung, and least popular brand is Motorola. Among three Web sites maximum brand-related positive discussions have been held on social networking site Facebook then on e-commerce site Flipkart, and minimum discussions held on review site CNET.

6.4 Result3: sentiment analysis-based brand performance evaluation

The analysis of how users react and respond to a product is as important as the analysis of content itself. Therefore, we discuss statistical comparison of sentiments of comments, review or discussion held across three domains for LG brand. In this experiment, we assigned a sentiment value of user reaction of all domain sites by computing the number of positive, negative and neutral over all the words under observation. In this work, our analysis is restricted to adjectives, as we have the highest accuracy of adjective words in used WordStat sentiment dictionary.

Our intuition is that terms used to form sentences may provoke strong reactions of approval or denial of sentence and determine the page rank on the basis of sentiment analysis.

We then analyzed the sentiment 'positive,' 'negative,' or 'neutral' for each sentence of networked web pages to assign sentiment-based page rank. Here in Fig. 8, we have shown results for LG brand phones for three chosen

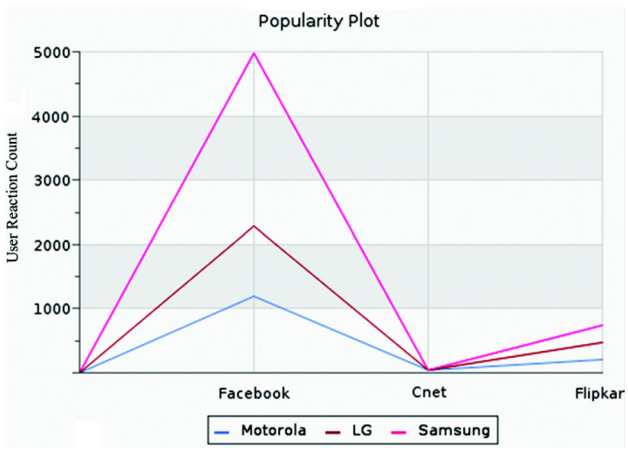

Fig. 7 Brand popularity comparison on Facebook, CNET and Flipkart

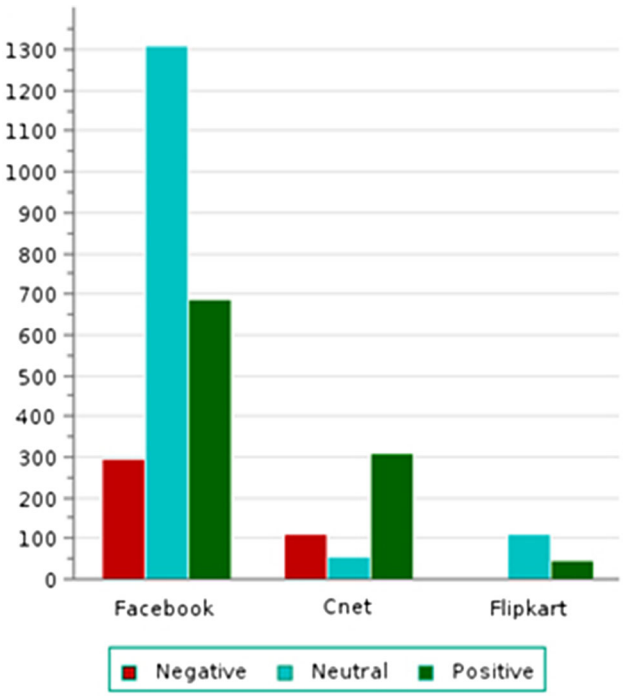

Fig. 8 Sentiment analysis for LG Brand on Facebook, CNET and Flipkart

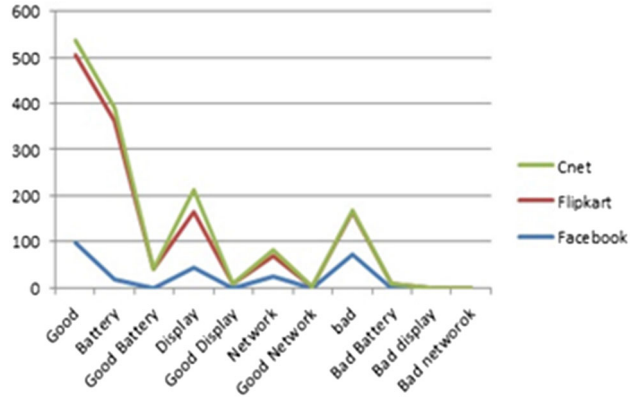

Fig. 9 Text edge processing for LG brand phones

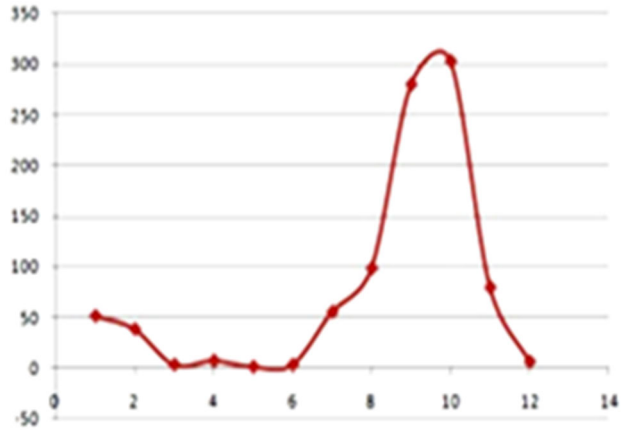

Fig. 10 Brand trend analysis for Motorola phone

domain Web sites and result shows Flipkart dataset observed pattern is substantially different and not having any negative sentiment for LG product and even does not have positive and neutral sentences too in high amount.

6.5 Result4: Text edge processing to analyze brand popularity with respect to brand features

In Fig. 8, illustrated embodiments are LG brand phone and shown graph represents number of relationship count between the main term of interest and found feature word in the three domain extracted sentences. As shown in Fig. 8, Flipkart doesn't have negative sentiments; same results have been achieved through test edge processing. CNET review site and Flipkart e-commerce site feature-based brand popularity graphs give almost the same popularity graph. Using text edge processing algorithm, we are able to analyze feature-based brand popularity (Fig. 9).

Using this approach, we analyzed brand trend. Brand trend analyzes brand feature review over a period of time using text edge processing and is able to check Web site popularity for any specific type of brand/product with respect to features. Brand trend with respect to feature is shown in Fig. 10 which represents the popularity of 'Motorola phone' over a certain period of months on Facebook.

7 Research contributions

Hence, our research work achieved contributions depicted by proposing a generalized brand analysis framework for online marketing are as follows-

- We proposed a brand popularity framework that incorporates the cumulative page rank with all the web page links. This page rank is a web page similarity measure with respect to search query. A web page having highest page rank is the most popular web page to advertise for that brand that was searched in the search query.
- We formed a network for resultant web pages with respect to search query searched on three search engines, and the network formed also covered related and recommended web pages of resultant search web pages.
- Assigned page rank to all networked web pages using social graph processing, text edge processing (Goeldi 2014) and sentiment analysis (Agarwal et al. 2011; Yessenov and Misailovic 2009; Pang et al. 2002).
- Analysis of three mobile phones brands—LG, Motorola and Samsung on three sites—social networking site Facebook, e-commerce site Flipkart and review analyze site CNET has been done to validate the performance of the work done.

8 Conclusion and future work

Due to the increasing use of Internet and the increasing demand for e-commerce sites, it becomes necessary for the companies to check their popularity, analyze the sentiments over the domain sites, and check the sites that have the maximum probability of being surfed when the user goes into surf for the brand.

So our methodology offers analysis of the sentiments of the users of the online social networking, analysis of the sentiments of the blog writers and analysis of the sentiments of the e-commerce users. Also our work offers a way in which the comments can be analyzed to find out which sites gain importance in terms of words like 'good display,' when these terms are used in conjunction with a product. Our project also offers a list of the Web sites which are most probable when the user enters a particular query in different search engines.

Although our present approach came up by exploring and refining several alternatives, we are well aware that it is by no means perfect or complete. Our future work will focus on the following topics:

- This approach at present works only for the category of mobile phones. Extending the project so that it becomes valid for all categories as well as brands and their products.
- The algorithm for sentiment analysis does not care whether the post has been written by an ordinary buyer, a random person or a proper blog writer who is followed by many people because the brand sentiment will change accordingly. For example, if a blog writer who is followed by many people criticizes a product while on the other hand the general user's opinion is positive about the product, then the sentiments for the brand would change accordingly.
- Checking the users' preference from their social networking sites and searching their history to check their e-commerce preferences and suggesting to the users products based on the above.
- Once the web pages are crawled through, checking the accuracy of the data on the web pages through algorithms like LDA, which could improve the existing algorithm of social graph processing.

References

Agarwal A, Xie B, Vovsha I, Rambow O, Passonneau R (2011) Sentiment analysis of twitter data. In: Proceedings of the workshop on languages in social media, Association for Computational Linguistics, pp 30–38

Arora A, Taneja V, Parashar S, Mishra A (2016) Cross-domain based event recommendation using tensor factorization. Open Computer Science 6(1):126–137

Bansal S, Gupta C, Arora A (2016) User tweets based genre prediction and movie recommendation using LSI and SVD. In: Contemporary computing (IC3), 2016 Ninth international conference on IEEE, pp 1–6

Behl D, Handa S, Arora A (2014) A bug mining tool to identify and analyze security bugs using Naive Bayes and tf-idf. In: Optimization, reliability, and information technology (ICROIT), 2014 international conference on IEEE, pp 294–299

Go A, Bhayani R, Huang L (2009) Twitter sentiment classification using distant supervision. CS224N Project report, Stanford, vol 1, p 12

Goeldi A (2014) Website network and advertisement analysis using analytic measurement of online social media content. U.S. Patent No. 7,974,983, 5 July 2011

Jamali S, Rangwala H (2009) Digging digg: Comment mining, popularity prediction, and social network analysis. In: Web information systems and mining, 2009. WISM 2009. International conference on IEEE, pp 32–38

Merriman DA, O'connor KJ (1999) Method of delivery, targeting, and measuring advertising over networks. U.S. Patent No. 5,948,061, 7 Sept 1999

Nigam K, Lafferty J, McCallum A (1999) Using maximum entropy for text classification. In: IJCAI-99 workshop on machine learning for information filtering, vol. 1, pp 61–67

Pang B, Lee L, Vaithyanathan S (2002) Thumbs up?: Sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on empirical methods in natural language processing, Association for Computational Linguistics, vol 10, pp 79–86

Siersdorfer S, Chelaru S, Pedro JS, Altingovde IS, Nejdl W (2014) Analyzing and mining comments and comment ratings on the social web. ACM Trans Web (TWEB) 8(3):17

Spencer J, Uchyigit G (2012) Sentimentor: Sentiment analysis of twitter data. In: Proceedings of European conference on machine learning and principles and practice of knowledge discovery in databases, pp 56–66

Tan S, Cheng X, Wang Y, Xu H (2009) Adapting naive bayes to domain adaptation for sentiment analysis. In: European Conference on Information Retrieval. Springer, Berlin Heidelberg, pp 337-349

Yessenov K, Misailovic S (2009) Sentiment analysis of movie review comments. Methodology 17:1–7