Distinguishing the Popularity Between Topics: A System for Up-to-date Opinion Retrieval and Mining in the Web

Nikolaos Pappas[1,2], Georgios Katsimpras[2] and Efstathios Stamatatos[2]

[1] Idiap Research Institute, Rue Marconi 19, 1920 Martigny, Switzerland
nikolaos.pappas@idiap.ch
http://www.idiap.ch/
[2] University of the Aegean, Dep. of Information and Communication Systems Engineering, Karlovassi 83200, Samos, Greece
gkatsimpras@gmail.com,stamatatos@aegean.gr
http://www.icsd.aegean.gr/

Abstract. The constantly increasing amount of opinionated texts found in the Web had a significant impact in the development of sentiment analysis. So far, the majority of the comparative studies in this field focus on analyzing fixed (offline) collections from certain domains, genres, or topics. In this paper, we present an online system for opinion mining and retrieval that is able to discover up-to-date web pages on given topics using focused crawling agents, extract opinionated textual parts from web pages, and estimate their polarity using opinion mining agents. The evaluation of the system on real-world case studies, demonstrates that is appropriate for opinion comparison between topics, since it provides useful indications on the popularity based on a relatively small amount of web pages. Moreover, it can produce genre-aware results of opinion retrieval, a valuable option for decision-makers.

Keywords: Opinion Retrieval, Text Mining, Sentiment Analysis, Information Extraction, Utility-Based Agents

1 Introduction

A huge number of user-generated content on various topics is created every day in social networks, news media, blogs, discussion forums and other sources in the Web. This content oftenly expresses opinions of users about certain products, people, services, etc. and therefore the need of computational treatment of opinion, sentiment, and subjectivity in text has become crucial [12]. Many applications, such as brand analysis, measuring marketing effectiveness, influence network analysis and many more, exploit the existing opinionated information.

During the last decade, considerable progress has been achieved in opinionated document retrieval. Most of the published studies are targeting blogs (TREC) [7, 10] and can be roughly categorized into two categories: lexicon-based [9, 21] and classification-based [4, 22]. The former utilize subjective dictionaries and decide whether the occurrences of these words suggest an opinionated document. The latter, develop subjectivity classifiers, based machine learning on opinionated and non-opinionated text. The proposed approaches are using fixed and offline collections of texts, taken from certain domains (e.g. blogs, movie reviews, message boards) or certain corpora.

In addition, opinion mining conclusions can differ according to the examined web genres (e.g. certain products may have good promotion articles but poor comments in blogs). So far, the task of collecting online domain-independent opinionated texts from various web sources in order to be used for opinion mining applications, has not been studied thoroughly. Moreover research on focused crawling usually deals with the more general task of collecting any kind of documents about a certain topic (e.g., [1,3,11]). However, opinion mining applications require the discovery of certain web genres that mostly comprise opinionated texts. Moreover, it is not yet possible to estimate the number of opinionated texts needed to extract reliable conclusions on the total polarity of opinions about particular topics.

In this paper, we present an online system for opinion retrieval and mining which handles the above subjects together: it discovers up-to-date topic-related documents dynamically from web sources using focused crawling techniques by targeting to specific genres (news, blogs, discussions) which are highly likely to contain opinionated texts; detects user-generated content regions inside the related pages by using web segmentation and noise removal techniques; computes a confidence score which quantifies the relatedness of the page to the given topic; and lastly performs automatic subjectivity and polarity detection on the sentences of the detected regions.

The main contribution of this paper is four-fold: (a) a unified framework for the discovery of topic-related opinionated texts in web pages, (b) a genre-based analysis of topic popularity[3], (c) a sentiment score estimation of opinionated regions of web pages, and (d) an efficient approach to estimate the sentiment polarity of topics using a few hundred documents.

The rest of this paper is organized as follows. Section 2 reviews the related research work. Sections 3, 4 and 5 provide an overview of the system and its components, whereas Section 6 describes the examined case studies. Finally, Section 7 summarizes the conclusions drawn from this study.

2 Related Work

There is a large body of research conducted for opinion retrieval and mining since TREC Blog was introduced in 2006 [10]. Most of these approaches are performing in a two-stage retrieval model. Firstly, one of the standard Information Retrieval methods is applied for locating topically relevant documents and secondly, various opinion mining/sentiment analysis algorithms are used to discover and identify opinionated texts within the documents.

The aforementioned approaches focus on detecting the subjectivity for each document, using various opinion mining methods such as subjectivity word/phrase

[3] We refer to popularity using the definition i.e. 'well-liked, admired by the people'. The detected positive and negative opinions of the people are used as indications for their admiration degree for a given topic.

dictionaries [9,20,21], machine learning algorithms [22] or proximity and phrase matching [19]. In [9], is presented a system which consists of three major modules: a fact-oriented information retrieval, dictionary-based opinion mining method and spam filtering. The information retrieval module in [20] utilizes proximity and phrase matching while the opinion module integrates a number of factors, such as frequency-based heuristics, special pronoun patterns and adjective/adverb-based heuristics. Zhang et al. [22] perform a concept-base information retrieval [5], machine learning opinion detection and a ranking algorithm for filtering the irrelevant information.

Many other related works utilize machine learning techniques such as SVMs [4] or focus on subjective/polarity classification [16–18]. In [4], SVM is used to classify sentences as opinionated or non opinionated, then decide whether the sentences are topic-specific and lastly compute a total document score by summing the SVM scores of the examined sentences. In [17], subjective language features are identified, such as low-frequency words, word collocations, adjectives and verbs, from corpora and used them in the subjectivity classification. In a more recent approach [2], Gelani et al. proposed a probabilistic model using proximity information of opinionated terms.

3 Overview of the System

The architecture of the proposed system is displayed in Fig. 1. The two major components are the Crawling Module and the Mining Module. The first is responsible for gathering relevant documents to a specific topic, while the second extracts and identifies opinionated documents. Both components are operating asynchronously using the Messaging Module to communicate[4], which provides scalability and robustness. The code for the system is available online[5].

Based on given topic query, the first task is to find a set of appropriate seed pages to guide the crawling procedure. To this end, the query is sent via Seeding Module to major search engines (e.g., Google, Yahoo, etc.) and the top results of each search engine, form the list of seed web pages. These results are stored in a distributed object memory and forwarded to the Crawling Module which initializes n Focused Crawler Agents (FCAs), each one using an equally-sized chunk of seeds while the crawled URLs are stored in a distributed database[6].

At the same time, n Opinion Miner Agents (OMAs) are initialized to process the web pages discovered by each FCA. The OMAs are responsible to segment the page into textual parts and filter out the non-informative parts (i.e., non-opinionated texts or texts irrelevant to the query) and then decide about the subjectivity and the polarity of each opinionated text.

4 Discovery of Topic-Related Web Documents

The information retrieval component of the system is a state-of-art focused crawling procedure. The idea is that, given a query, up-to-date relevant documents

[4] http://www.rabbitmq.com/

[5] https://github.com/nik0spapp/icrawler

[6] http://www.mongodb.org/

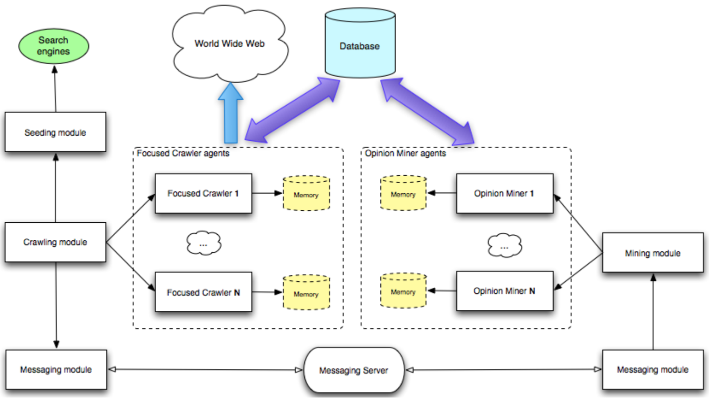


Fig. 1. The basic architecture and the components of the system.

can be retrieved from various domains and web-genres by following the path of a focused crawler, but also in a real-time manner. For the purposes of our system, [13] is especially suitable. It is an agent-based focused crawling framework that is able to retrieve topic- and genre-related web documents in an automated and real-time manner.

The focused crawler agents displayed in Fig. 1 are making use of a utility function that weights an unvisited URL p and consists of two components: one for the topic relevance and one for the genre relevance.

$$Linkscore(p) = w_T * Linkscore_T(p) + w_G * Linkscore_G(p) \tag{1}$$

The $Linkscore_T$ and $Linkscore_G$ are relevance scores based on topic and genre accordingly; and they are computed by using link analysis techniques (see [13]). For our experiments we used equally weighted these two scores ($w_T = w_G = 0.5$) since it has been shown that it leads to both topic and genre related document discovery. In addition, we selected the news, blogs and discussions genres for seed URLs and for weighting the genre component in the above equation, since these genres are more likely to contain opinionated texts. For the implementation we used Scrapy, a python-based crawling framework[7].

5 Opinion Retrieval and Mining

The Mining Module is responsible for the extraction of the opinionated textual parts from web pages and the estimation of their sentiment polarity. An OMA performs web page segmentation, assigns a confidence score which indicated the relevancy of the document being processed and estimates the sentiment subjectivity and polarity of the page. It learns from its previous experience with a page and uses this knowledge for solving more accurately and the sentiment analysis problem in future processing (Section 5.3).

In Fig. 2 the page processing by the OMA is displayed. Initially, it receives a message from an FCA to perform a task, connects to the corresponding database and retrieves all the relevant pages. Then, for each page, three basic procedures are executed; web page segmentation, page filtering, and sentiment analysis.

[7] http://scrapy.org/

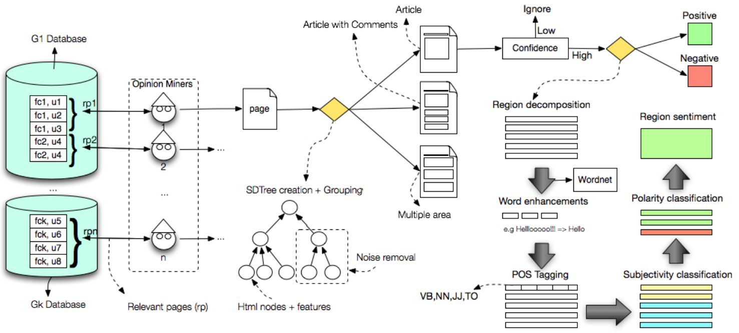
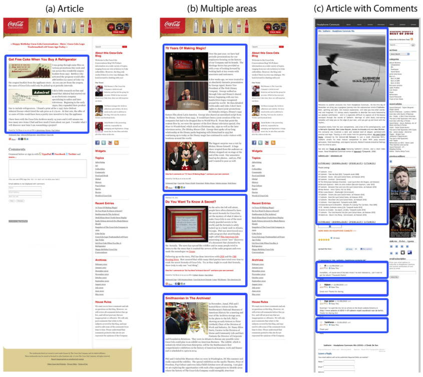


Fig. 2. The processing steps of an OMA are displayed for a given page: (a) web page segmentation (b) page filtering and (c) sentiment analysis.

5.1 Web Page Segmentation

For this task a mechanism is needed to segment a web page into semantically-coherent parts that correspond to the basic textual components of the web page. Moreover, it is convenient that the noisy segments (i.e., ads, banners, etc.) are removed. A very recent approach that handles the above issues in an efficient manner, is presented in [14]. It exploits visual and non-visual characteristics of a web page encapsulated in a DOM Tree with additional features, called SD-Tree, and performs the layout classification and extraction using SD-algorithm.

We adopted this method because it provides robust identification of informative textual parts and it yields promising results as a web page type classifier in a realistic web setting. The output of this processing is a set of informative annotated regions in to three possible classes (Article, Multiple areas and Article with comments). Output examples are displayed in Fig. 3.


(a) Article

(b) Multiple areas

(c) Article with Comments

Fig. 3. Example outputs of the SD algorithm for the three possible classes: (a) Article, (b) Multiple areas and (c) Article with comments.

5.2 Page Filtering

The web page segmentation mechanism provides a set of segments with informative text of user-generated content; a source of potential opinions. However, it is not yet clear whether each extracted segment refers to the given query or another subject. There is a chance that the existence of the query in the document at the retrieval stage was not present on the informative regions (e.g. it was part of the ads). Therefore, we need a mechanism to filter out all the irrelevant pages by assigning *confidence score* to each detected region and by filtering out pages with low score (i.e. unlikely to refer to the given query).

The *confidence score* for a page i is calculated by the weighted combination of the presence of the topic in the detected regions, the URL and the title:

$$\begin{aligned} Confidence_i = {} & w_1 * ArticleContextScore(i) \\ & + w_2 * CommentsContextScore(i) \\ & + w_3 * MultipleContextScore(i) \\ & + w_4 * UrlScore(i) \\ & + w_5 * TitleScore(i) \end{aligned}$$

Regarding the type of the document, some context scores of the above formulation may be equal to 0. The weights can be learned from an annotated corpus of region class and relevance value pairs. For our experiments we used the weights below which yielded good results for each of the classes: (a) Article: ($w_1 = 0.4$, $w_4 = 0.3$, $w_5 = 0.3$), (b) Article with comments: ($w_1 = 0.2$, $w_2 = 0.2$, $w_4 = 0.2$, $w_5 = 0.4$) and (c) Multiple areas: ($w_3 = 0.4$, $w_4 = 0.3$, $w_5 = 0.3$). For example, given the query *Audi*, if it is present in the title and URL the confidence would be: $0.4 * 0 + 0.3 * 1 + 0.3 * 1 = 0.6$.

In the case some of the non-zero weighted regions are missing from the page, their weights are distributed equally to the rest of the coefficients. To this end, we select the documents with high confidence scores based on a threshold t. The threshold values range from 0 to 1. The closer to 1 the threshold is, the greater the confidence about the topic. For the experiments we used the value of $t = 0.6$.

5.3 Sentiment Analysis

The confidence mechanism provides related documents to a given topic. The next step is to to detect whether a given document contains subjective information or not. In order to learn dynamically the domain knowledge for a given query we use self-trained machine learning algorithms (see [6,15]). Initially, the filtered regions are decomposed into sentences (Fig. 2). The sentences are then pre-processed in three steps: (a) tokenization, (b) spell-checking based on WordNet and (c) part-of-speech (POS) tagging. Next, the set of sentences in the text area is given as input to our subjectivity classifier. Each sentence is classified as subjective or not. All sentences that are labeled as subjective are then forwarded to our polarity classifier. And thus, the sentiment for each sentence is determined.

Subjectivity classification We adopted the method presented in [15] which is a bootstrapping process that learns linguistically rich extraction patterns for subjective expressions. High-precision classifiers using a subjectivity lexicon (MPQA[8]), label unannotated data to create a large training set, which is given to an extraction pattern learning algorithm. The learned patterns are then used

[8] http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/

to identify more subjective sentences. The bootstrapping process learns many subjective patterns and increases recall while maintaining high precision. To make the learning algorithm tractable in an online setting, we activate only the n-most frequent patterns at each learning step.

Polarity classification Similarly, we adopted a bootstrapping method presented in [6]. The method follows three steps: (a) rule-based polarity classification with high precision [18], (b) training of an SVM classifier[9] using as input data the high scored instances from the rule-based classifier and (c) classification with the self-trained SVM classifier. The rule-based polarity classifier makes use of a subjectivity lexicon (MPQA[4]) and proceeds as follows: preprocessing, feature extraction, polar expression marking, negation modeling, intensifier marking, heuristic weighting and classification. Since we target on web text, we further extended the MPQA lexicon with informal and swear words as well as a great amount of emoticons. Lastly, for tractability reasons, we trained the SVM for a given query in a first short run and then we use it online in a second longer run.

Total Sentiment Estimation Let D be a set of topic-related documents, r_{ij} the i-th region of document dj, and $Score(r_{ij})$ the sentiment score of r_{ij}. Then, the total *sentiment score* is defined as follows:

$$TotalScore(D) = \sum_{d_j \in D} \left(\sum_{r_{ij} \in d_j} Score(r_{ij}) \right) \in R \qquad (2)$$

Unlike the Eq. 2 where the detected regions are treated equally, the *normalized sentiment score* weighs them based on the region length as follows:

$$NormalizedScore(D) = \sum_{d_j \in D} \left(\sum_{r_i \in d_j} \frac{Score(r_{ij})}{|r_{ij}|} \right) \in R \qquad (3)$$

where $|r_{ij}|$ is the length of the region r_{ij} in words. Lastly, given a the set of regions with positive sentiment score r_{pos} in D and r_{neg} with negative sentiment score accordingly, we compute the *sentiment ratio* as follows:

$$SentimentRatio(D) = \frac{|r_{pos}|}{|r_{pos}| + |r_{neg}|} \in [0,1] \qquad (4)$$

6 Experiments

In this section we examine the overall effectiveness of the proposed system to estimate the total sentiment polarity of the retrieved opinions for a given topic query in the Web. The study focuses on the system's ability to provide structured sentiment analysis results as well as on the number of pages required to form a reliable calculation of the sentiment. Since the system is designed to run in the Web (web pages not yet necessarily indexed by search engines), it is more appropriate to evaluate in real-world case studies rather than offline collections. The selected case studies concern well-known subjects that enable us to properly validate the produced results and were performed in October 2011.

[9] http://pyml.sourceforge.net/tutorial.html#svms

6.1 Case Study 1: Distinguishing the Popularity Between Topics

In the first case study we examined queries on two well-known political concepts: *democracy* and *fascism*. The presented system was used to discover a predefined number of relevant web pages for each query (1,000 relevant pages), extract the opinionated texts from them and calculate their sentiment polarity.

Figure 4 depicts the distribution of the detected relevant text regions over three major types (articles, multiple areas, and comments), the total sentiment score, and the total normalized sentiment score for both queries. As expected, the *democracy* query has a far more positive sentiment score in all three region types. In addition, the *fascism* query has negative sentiment scores in two types of pages (articles and multiple areas). Interestingly, the relatively high sentiment score of the *fascism* query for comments indicates an increased use from people with far-right radical political opinions.

The normalized sentiment score seems to be able to better represent the differences in sentiment polarity since it takes into account the length of the extracted text regions. For example, in articles usually there are a lot of long sentences with neutral polarity so the overall sentiment score tends to be lower. On the other hand, the normalized sentiment score indicates the intensity of the positive or the negative sentiment polarity.

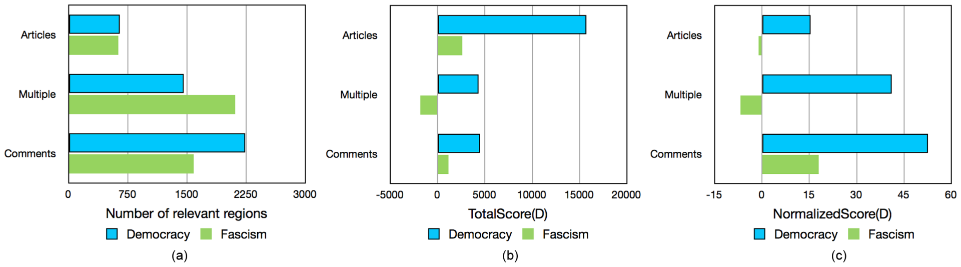


Fig. 4. Overall results for *democracy* and *fascism*: (a) number of relevant regions, (b) total sentiment scores and (c) total normalized sentiment scores per region type.

A more detailed look in the distribution of sentiment polarity with respect to the three region types is given in Fig. 5 for *democracy* and *fascism* queries. In the former case, the positive sentiment is dominant in all region types with more emphasis in articles. Despite the increased percentage of neutral polarity in multiple areas and in comments, the positive opinions are in all cases greater than the negative opinions with an average difference of 20%. In the latter case, the negative polarity is greater than the positive one in most of the regions (articles and multiple areas). The difference of the positive versus negative polarity is not so intense in the comment regions.

Finally, Fig. 6 shows the sentiment ratio (Eq. 4) for both queries (y-axis) during the process of discovering relevant web pages (x-axis) on these topics. The sentiment ratio remains practically stable after a few hundred pages have been examined. Moreover, there is a notable difference between the sentiment scores of the two queries indicating a much more positive polarity for *democracy* in comparison to *fascism*. This means that we can reliably decide about the sentiment polarity in short time.

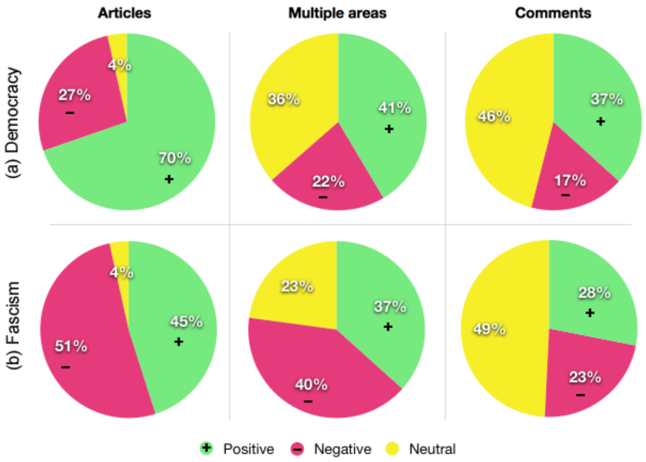


Fig. 5. Percentage of sentiments per region type for (a) *democracy* and (b) *fascism*.

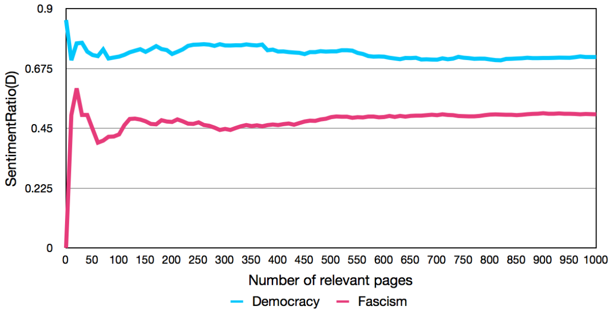


Fig. 6. Sentiment ratio curves for *democracy* and *fascism* queries.

6.2 Case Study 2: Ranking of Competitive Products

This experiment focuses on the examination of the system when it deals with a set of queries on competitive products in the same thematic area. In this case, it is crucial to provide comparative sentiment results and decide about a general ranking of the products according to the opinions found on web pages. We used a threshold of 300 relevant pages to be discovered for each of the product queries in the set. Given the same number of relevant pages the products can be compared based on the total sentiment estimations of the detected region types and the discovered pages overall (Eq. 2, 3).

Soft drinks Five well-known soft drinks were used as queries: Pepsi, Dr. Pepper, Sprite, 7up and Fanta. Figure 7 shows the topic-related region types, the total sentiment estimation scores per region type. Based on the distribution of the detected region types, Pepsi and Dr. Pepper are more frequently discussed in multiple areas (usually blogs, forums) and article with comments.

The total sentiment scores have similar values for most of the soft drinks and sentiment distinction is not very clear. A closer look reveals that 7up, Sprite, and Fanta have a particularly high score in pages with articles, potentially the result of promotion. Conversely, the normalized sentiment score highlights the differences between the products more clearly; it gives greater emphasis to pages with multiple opinionated areas and provides a different aspect in the evaluation of opinions (potentially of end users) about the products.

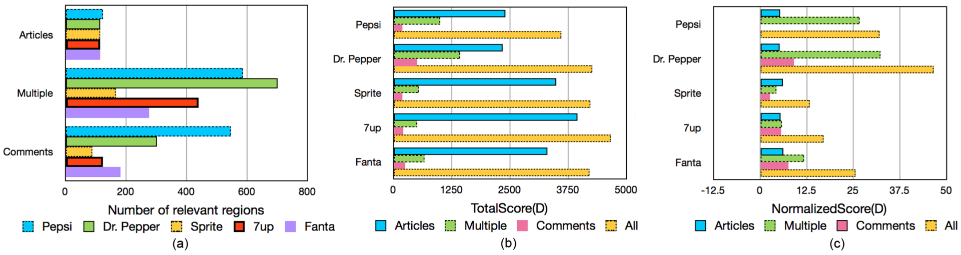


Fig. 7. Overall results for soft drinks: (a) number of relevant regions per type, (b) total sentiment score and (c) total normalized sentiment score per soft drink.

Lastly, we compared the ranking based on the total sentiment estimation (Eq. 2, 3) to the ranking of the soft drinks based on social media metrics (number of likes, number of people talking) of their major groups on Facebook (Table 1). The ranking based on the normalized sentiment score matches closely to the one obtained based on the social media metrics. Sprite is probably low ranked due to the neutral or negative opinions found. Also, it has the smallest number of talking people from all the soft drinks.

Table 1. List of soft drinks and IM clients ranked by the social media metrics and the rankings based on total sentiment score and total normalized sentiment score.

Rank	**Soft drink**	**Likes**	**Talking**	**Both**	***TotalScore***	***NormalizedScore***
1_{st}	Dr. Pepper	12,093,912	187,011	12,280,923	7up	Dr. Pepper
2_{nd}	Pepsi	11,835,244	236,105	12,071,349	Dr. Pepper	Pepsi
3_{rd}	Sprite	8,574,563	50,192	8,624,755	Sprite	Fanta
4_{th}	Fanta	2,650,072	84,080	2,734,152	Fanta	7up
5_{th}	7up	785,967	75,996	861,963	Pepsi	Sprite
	IM Client	**Followers**	-	-	***TotalScore***	***NormalizedScore***
1_{st}	Google Talk	405,818	-	-	Google Talk	Google Talk
2_{nd}	Skype	367,385	-	-	Skype	Skype
3_{rd}	MSN	82,896	-	-	MSN	MSN
4_{th}	AOL	14,431	-	-	AOL	ICQ
5_{th}	ICQ	14,138	-	-	ICQ	AOL
				NDCG:	0.841	0.993

Instant Messaging (IM) clients Similarly, some well-known IM clients were also used: Google talk, Skype, MSN messenger, AOL messenger and ICQ. We compared the ranking based on the total sentiment estimation (Eq. 2, 3) to the ranking of them based on their followers in Twitter[10]. In this case, the ranking based on each of the estimation scores matched almost perfectly the ranking based on the social media metrics. AOL and ICQ were ranked falsely based on

[10] Some of the IM clients' official groups were missing from Facebook (e.g. Google talk).

the normalized score but they were not clearly distinguishable either based on the number of followers (14,431 and 14,138 accordingly).

Finally, we computed the average normalized cumulative gain (NDCG) [8] for both soft drinks and IM clients. In Table 1, we can observe that the normalized sentiment score performed better than the simple one in the examined queries. The long subjective sentences seem to be less important than shorter ones in the total estimation over the text regions.

7 Conclusions and Future Work

We presented an online system for topic-based opinion retrieval and mining in the Web. Rather than making use of static well-defined document collections, we acquire dynamic collections in real-time from the Web. Such collections targeted to certain web genres, can provide up-to-date sources of opinionated text about a given topic. The opinion mining agents are able to extract opinionated textual parts from web pages and estimate their sentiment polarity while ignoring irrelevant and noisy regions. Useful conclusions can then be drawn based on the distribution of positive and negative opinions over the detected regions.

A series of experiments demonstrated that the system can provide a total estimation about the popularity of certain topics as well as comparative results for competitive topics. The genre-aware output of the sentiment results, can be of crucial importance for decision-makers since they can estimate the result of promotion as well as the potential difference in the opinion between the general population and some influential people. In addition, the system provides efficient results since a few hundred web pages are usually enough to estimate the total sentiment polarity about a given query.

A dimension of the system that could be further explored concerns the date that each opinionated text was created. This temporal information can be used to express the change of sentiment polarity about a certain topic over time and to provide an in-depth analysis for a certain time period.

8 Acknowledgements

The work described in this article was supported by the European Union through the inEvent project FP7-ICT n. 287872 (see http://www.inevent-project.eu).

References

1. X. Chen and X. Zhang. Hawk: A focused crawler with content and link analysis. In *Proc. of the International Conference on e-Business Engineering (ICEBE)*, Xi'an Jiaotong Xian, China, 2008.
2. S. Gerani, M. J. Carman, and F. Crestani. Proximity-based opinion retrieval. In *33rd International Conference on Research and Development in Information Retrieval (SIGIR)*, Geneva, Switzerland, 2010.
3. D. Hati, B. Sahoo, and A. Kumar. Adaptive focused crawling based on link analysis. In *Proc. of 2nd International Conference on Education Technology and Computer (ICETC)*, Shanghai, China, 2010.
4. L. Jia, C. Yu, and W. Zhang. Uic at trec 2008 blog track. In *Proc. of The 17th Text Retrieval Conference (TREC)*, Gaithersburg, USA, 2008.

5. S. Liu, F. Liu, C. Yu, and W. Meng. An effective approach to document retrieval via utilizing wordnet and recognizing phrases. In *Proc. of the 27th International Conference on Research and Development in Information Retrieval (SIGIR)*, Sheffield, United Kingdom, 2004.
6. D. K. M Wiegand. Bootstrapping supervised machine-learning polarity classifiers with rule-based classification. In *Proceedings of the 1st ECAI-Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA)*, Lisbon, Portugal, 2009.
7. C. Macdonald, I. Ounis, and I. Soboroff. Overview of trec-2009 blog track. In *Proc. of The 17th Text Retrieval Conference (TREC)*, Gaithersburg, USA, 2009.
8. C. D. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval.* Cambridge University Press, 2008.
9. G. Mishne. Multiple ranking strategies for opinion retrieval in blogs. In *Proc. of the 15th Text Retrieval Conference (TREC)*, Gaithersburg, USA, 2006.
10. I. Ounis, M. de Rijke, C. Macdonald, G. Mishne, and I. Soboroff. Overview of the trec-2006 blog track. In *Proc. of TREC*, Gaithersburg, USA, 2006.
11. A. Pal, D. S. Tomar, and S. C. Shrivastava. Effective Focused Crawling Based on Content and Link Structure Analysis. *Journal of Computer Science*, 2(1), 2009.
12. B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.
13. N. Pappas, G. Katsimpras, and E. Stamatatos. An agent-based focused crawling framework for topic- and genre-related web document discovery. In *Proc. of the 24th IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, Athens, Greece, 2012.
14. N. Pappas, G. Katsimpras, and E. Stamatatos. Extracting informative textual parts from web pages containing user-generated content. In *Proc. of the 12th International Conference on Knowledge Management and Knowledge Technologies (i-Know)*, Graz, Austria, 2012.
15. E. Riloff and J. Wiebe. Learning extraction patterns for subjective expressions. In *Proc. of the International Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Sapporo, Japan, 2003.
16. P. D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proc. of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*, Philadelphia, USA, 2002.
17. J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin. Learning subjective language. *Computational Linguistics*, 30(3):277–308, 2004.
18. T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *International Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP)*, Vancouver, Canada, 2005.
19. K. Yang. Widit in trec 2008 blog track: Leveraging multiple sources of opinion evidence. In *Proc. of The 17th Text Retrieval Conference (TREC)*, Gaithersburg, USA, 2009.
20. K. Yang, N. Yu, R. Valerio, and H. Zhang. Widit in trec-2006 blog track. In *Proc. of The 14th Text Retrieval Conference (TREC)*, Gaithersburg, USA, 2006.
21. M. Zhang and X. Ye. A generation model to unify topic relevance and lexicon-based sentiment for opinion retrieval. In *Proc. of the 31st International Conference on Research and Development in Information Retrieval (SIGIR)*, Singapore, 2008.
22. W. Zhang, C. Yu, and W. Meng. Opinion retrieval from blogs. In *Proc. of the 16th International Conference on Information and Knowledge Management (CIKM)*, Lisbon, Portugal, 2007.