A Novel Approach for Social Behavior Analysis of the Blogosphere

Reza Zafarani, Mohammad-Amin Jashki, Hamidreza Baghi, and Ali A. Ghorbani

Faculty of Computer Science, University of New Brunswick
Fredericton, NB, Canada
{r.zafarani,a.jashki,hamid.baghi,ghorbani}@unb.ca

Abstract. The web, as a real mass medium, has become an invaluable data source for Information Extraction and Retrieval systems. Digital authoring is a relatively new style of communication, usually facilitated by computer networks and the Internet. We believe that the behavior of the people in cyberspace can be a representative of the real social behaviors and that this data can be employed to analyze the behavior of a society. In this paper we have used blogs as the main representative of this digital data. A system of blog analyzing, named *Blogizer*, has been designed to analyze these blogs. The system employs two specific measurements to determine the level of citizen engagement. The detailed analysis and the proof of concept case study provides promising results. Based on the obtained results, more than 70.52% of the topic assignments and 58.10% of the significance assignments were ascribed successfully.[1]

1 Introduction

Investigating social behavior has shown that sometimes some sensitive topics emerge as issues that can result in unexpected social consequences. For example, in an election, the results of the election might become unexpected in comparison to pre-election polls. The issues related to the election rise in pitch very quickly and within a short period of time become the key issues that result in a changing of the government. These issues might come onto the political/social stage only in the last few weeks of the election. However, there might be a way to mine the information in cyber discussion forums, the governments sources, local blogs, network payloads and other collaborative vehicles for evidence that these issues are emerging. Technology might also be used to detect the point of deflection when simple discussion becomes a burning political issue. The research in this paper is dedicated to provide the possible means to automate the process of this detection and data analysis.

To facilitate the discovery of socio-political issues, we need to first identify the possible topics (categories) of discussion. Then we need to employ some measuring techniques to determine the level of citizen engagement for each topic.

[1] The authors have had the same amount of contribution.

S. Bergler (Ed.): Canadian AI 2008, LNAI 5032, pp. 356–367, 2008.

This will give us a list of the possible topics and an indication of how strongly the citizenry feels on each of these topics.

In this paper we have used blogs as the main source of data. A blog analyzing system, named *Blogizer*[2], has been also designed to analyze these blogs. The system conducts two specific measurements to determine the level of citizen engagement which will be described in the following sections.

In the next section, we review the structure of the Blogizer system and how it is used to analyze the social behavior of blogs. In Sections 3-6, a detailed overview of different components of Blogizer system along with the methods used in each is given. Finally, the last two sections conclude with the attained simulation results, a final system analysis, and a detailed case study of our system.

2 The Blogizer System

The proposed model for the social behavior analysis of blogs has five components: namely, *corpus construction*, *preprocessing and vectorization*, *topic discovery*, *measurements*, and the *final analysis* component. These components along with the dataflow between them are depicted in Figure 1.

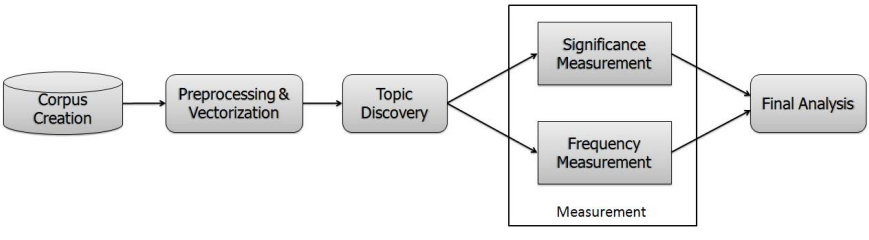

Fig. 1. The Blogizer Big Picture

In order to analyze topics for the emergence of political issues, a corpus of blog data is required. This corpus must be created using the available blog posts on the web and should be regularly updated. The *corpus creation* component takes care of this task in the Blogizer system. The created textual corpus is processed and its words are extracted, stemmed, and the stop words are eliminated. Finally, the posts are vectorized based on the remaining stemmed wordlist. The mentioned keyword extraction, stemming, and stopword removal takes place in the *preprocessing and vectorization* component.

In order to discover the possible topics, a mechanism for dynamic categorization of these vectorized posts is employed by the *topic discovery* subsystem. These extracted topics are the popular subjects discussed in the blog posts.

Following this phase, a set of measurements are conducted for two different metrics: *frequency* and *significance*. Frequency of a topic is the volume of discussion on that topic compared to the total volume of discussion. The temporal

[2] Blogizer stands for Blog Analyzer.

frequency of a specific topic is determined as a part of frequency measurement process.

Significance, also known as quality, is the measure of textual data worthiness and trustablitiy. It can be utilized in detecting worthwhile discussions in Information Retrieval systems. Eliciting Significance value in textual data is therefore to some extent a linguistic problem. Algorithms are required to determine the Significance of a text. Let us consider the following two sample sentences in order to assist in clarifying the significance factor:

> A message containing "*I don't trust you mister politician*" should be rated higher in significance if the message said "*I don't trust you mister politician because you promised to eliminate poverty and you didn't deliver.*"

Finally, the system provides a detailed analysis of the Blogs. This analysis can be extensively used to detect *issues*. Issues are the items of debate that require socio-political attention, and without that attention, citizens (or customers) will be unsatisfied and create difficulty for the government and the society.

3 Corpus Construction

The first stage in corpus construction is blog list gathering. As already mentioned, Blogizer uses blog data as its input. There exists lots of blogging websites on the Internet in numerous languages. However, a complete social/political repository of the blog and blogpost URLs is not currently available on the web[3]. Therefore, we designed a bot to extract blog links. The bot used a dictionary of around 1500 common words and constructed a 30-word random query from this dictionary in each execution time. This query was sent to Google Blog Search[4] and the retrieved blogs, which were identified by their URL, were saved. Only the top 20 related links to the search query were fetched. For the initial corpus creation we gathered a total number of 15326 blog links using this technique.

Furthermore, for the purpose of Social Network Analysis (SNA) and considering the relationship between different blogs, all friends of a given blog were extracted. A friend of a blog is a blog which is listed on the friends panel of that blog. The friend panel of a blog shows the blogs that are of interest to the authors of that blog. To extract this information, BlogRolling.com was used. This website provides a friend-list for each given blog. A bot was designed to automate this process to the retrieve this blog relationships. After this phase, the total number of blogs in our initial corpus was increased to 17548.

A crawler was designed to retrieve all posts in each given blog for later use. Around 280000 posts shared between 17548 blogs was retrieved in the initial executions of this crawler. Blogizer is dedicated to the analysis of English blogs; therefore, non-English blogs were filtered out after this stage. Language detection

[3] TREC BLOG06[1] Web corpus can be used for general purposes.
[4] http://blogsearch.google.com/

is performed using the statistical properties of the languages [2]. Non-English blogs can introduce noise in these systems due to the excessive impact they would have on the TF-IDF based phrase extraction phase, which will be explained in the following sections.

4 Preprocessing and Vectorization

Like any other text mining system, a conversion from unstructured text to structured data is required. For each post, the plain text is extracted and HTML tags are removed. The extracted text is tokenized; then, the stopwords are removed. The resulting tokens are stemmed. WordNet [3] and Porter [4] stemmers were used and Porter stemmer was finally selected because of its high performance. The documents are vectorized using TF-IDF and the wordlist of stemmed words. During vectorization, two filtering tasks are also performed. These tasks are described below.

Word List Creation and Pruning. A wordlist is generated from all given posts right after the stemming phase. This wordlist is pruned based on the frequency of the words to remove words with excessive usage (e.g. articles) or rarely used words (words with low occurence frequency).

Keyword Extraction. In order to extract representative keywords for each given post, the top 7 frequent words in each post based on the TF-IDF measure is selected as representative keywords of that document in the Topic Discovery phase.

5 Topic Discovery

Measurements are performed on posts in each topic; therefore, topics should be discovered in the given posts before this stage. There exists a variety of solutions to the topic discovery. Clustering of vectorized documents, is our provided solution to topic discovery. The measurement variations over time for each cluster will be analyzed by the system at the end.

For the clustering algorithm, different clustering methods were tested and evaluated. Among these, we concentrated more on K-Means, X-Means[5], and Bisecting K-Means [6]. The best results, as empirically proved by Steinbach et. al[6], were obtained by utilizing Bisecting K-Means.

Cluster Representatives. The clustering algorithm prototypes (centroids in here) are used to build cluster representatives. The centroid is typically not presented among the cluster documents; therefore, top 7 words from the closest document to the cluster centroid, based on the TF-IDF measure, are selected as the cluster representative. Table 1 contains cluster representatives for a number of clusters.

Table 1. Cluster Representatives for a Sample of 5 Clusters

john father church christian bishop pope doctrin
pound weight diet loss exercis lose calori
onlin internet medic engin search pharmaci medicin
congress democrat iraq troop bush presid petraeu
mississippi hurt happen nick bridg collaps sadden

6 Measurements

6.1 Frequency Measurement in Textual Data

Frequency measurement is one of the parts of the Blogizer measuring component. It provides the system with the necessary means to track the occurrences of events in a specific topic. Within each topic, the fraction of posts sent daily within that topic over the total number of daily posts in all the tracked topics represents the daily frequency of the topic. This value is saved for different topics and its variations are analyzed by the system through out the time.

6.2 Significance Measurement in Textual Data

Significance measurement is the other measurement part of the Blogizer measuring component. The process of significance measurement is shown in Figure 2. Significance measurement in Blogizer is performed based on two criteria: namely, *Significance measurement based on social network analysis* and *Content-based significance Measurement*.

Significance measurement based on Social Network Analysis. The impact of bloggers on each other plays an important role in measuring significance

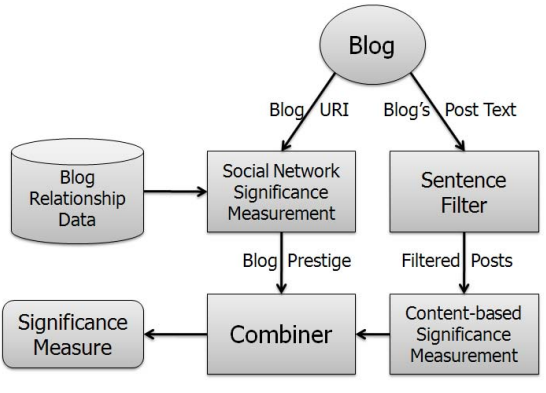

Fig. 2. The Significance Measurement Model

in measuring significance. The general assumption that those blogs which contain more significant posts are more attractive to other bloggers is made. This leads to the basic idea of employing Social Network Analysis algorithms, such as PageRank [7], to assign a value to each blog. This value, which we call the *blog prestige*, is one of the two final features used to measure the posts significance.

In order to detect the significance of the posts, the PageRank value of all blogs are calculated. Based on the friend-URL list gathered, which is the list of friends of a particular blog, a link graph is constructed. This graph is used by the PageRank algorithm to compute the PageRank values of the blogs. These computed values are later combined with the Content-based significance in the final signifance calculation.

Content-Based Significance Extraction. Significance can also be analyzed based on the textual content of the blog posts. The relevance between post data and the summaries created for blogs has been used as a measure in the system to quantify what we call content-based significance. Hence, a blog post related to a certain topic is more significant if its summary is closer to the summaries created for that specific topic. Let us consider the example of a news blog. The more this blog is containing the daily headlines the higher significance value should be ascribed to it. Therefore, to achieve this goal the following procedure of significance measurement is defined:

1. *Sentence Filter*: All the sentences in a given post are first fed into a significant sentence filtering which filters out sentences with low significance values or sentences which are unlikely to be significant.
2. *Post Summarization*: For each post a summary is created using the methods which will be discussed later in this section. This summary is in fact a good representation of what the post is talking about.
3. *Topic Summarization*: For each cluster of documents (posts on the same topic), a summary of all the posts is created. This summary is in fact a representative of the topic.
4. *Significance Measurement*: For a given post in a topic, a value is assigned to that post based on the closeness of its summary to the summary of the topic.

We have used an LSI based summarization [8] to create summaries for both posts and topics. The significant sentence filtering is inspired by the work done in [9]. This filter consists of five main components, which are demonstrated in Figure 3. These components are the pre-processor, POS-tagger, frequent pattern extractor, significance-signs and length extractor, and the classifier. Besides the normal pre-processing and POS-tagging, three different feature values are measured for each sentence: sentence length, number of occurrences of significance-demonstrative words, and the number of occurrences of a set of frequent grammatical patterns. Significance-demonstrative keywords are mostly English conjunctions which result in high significance sentences if present. Frequent patterns are grammatical patterns that are common in sentences with high

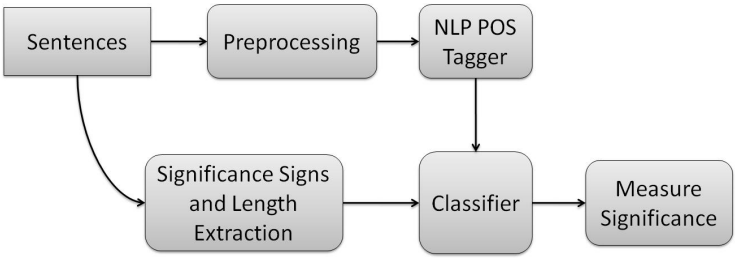

Fig. 3. The Architecture of the Proposed Model

Table 2. Frequent POS Tag patterns in sentences with high significance

IN PRP$	IN DT NN	DT NN VBZ TO VB	VBZ DT IN
DT NN IN DT JJ NN	PRP VBP	JJ NN IN IN IN JJ	
PRP VBD NN VBZ VBN NN IN DT	MD VB	VBZ DT JJ	

significance. These calculated feature values for different sentences are given to a SVM classifier which filters out the insignificant sentences.

A bag of 105 words is used as the set of significance demonstrative words. For example, the existence of a word such as *because* represents a sentence with a probable high degree of significance (see examples in introduction). A subset of used words is shown in Table 3.

Moreover, a set of frequent patterns is considered as a feature in our system. Apriori [10] frequent pattern mining algorithm is used to extract frequent patterns in hand-labeled POS-tagged sentences of high significance. In addition, a set of frequent patterns in a dataset of low significance sentences were extracted. Any frequent pattern in the high significance sentences which was also frequent in the low significance sentences was removed. This is to make sure that the set of patterns represent significant sentences. The number of existing patterns in a sentence is considered an important feature for the significance. A subset of patterns used is shown in Table 2.

From table 2 it can be seen that the frequent patterns are a set of consecutive grammatical roles represented by POS-tags which are to be matched with the POS-tagged input sentences in order to recognize high significance sentences. Different sentences may result in similar POS-tags as long as they obey a similar grammatical pattern; therefore, extracting frequent patterns from POS-tagged sentences reveals the underlying structure of significant sentences.

These features are calculated for each given sentence and the attribute values are provided to our binary classifier, which filters out improbable significant sentences.

The classifier was trained using sentences from three different datasets. The first dataset (generated by the authors) is gathered from conclusion sentences in scientific papers, which probably have high significance factor. The other sentences

Table 3. A Subset of the Bag of Words Used

for	and	nor	since	though	before
or	yet	so	unless	now that	when
because	although	if	while	as	whereas
in order that	until	in case	but	after	thus

were randomly selected from MPQA [11] and Subjectivity [12] datasets. All these sentences were hand-labeled. A total of 1000 sentences were hand-labeled and used in our simulation.

Significance Combiner. Currently, a combination of content-based significance value and the blog prestige value is used to calculate the final significance measure. This simplified version of significance measure is calculated by employing a polynomial utility function.

7 Results and Discussions

7.1 Topic Discovery Simulation Results

In order to evaluate the topic discovery algorithm, we selected a number of posts randomly and checked whether they were placed in a topic which was representative of the post. Two hundred random posts were selected from the post collection. For each post, human subjects determined if the topic, which is a set of seven representative words, corresponds to the subject of that post. Each document was shown to each human subject along with the seven keywords of the topic. Topic assignments were considered valid if a majority of human subjects approved the topic assignment.

The results showed that among these 200 posts, 122 were assigned correctly; this means that 61% of the posts were assigned to correct topics. Further analysis showed that among the topics, two containing noisy data existed. These two topics contain the posts which either are unrelated to the other topics or contain only pictures and videos. By removing these two topics and their related posts the accuracy is increased to 70.52%.

7.2 Significance Simulation Results

In the beginning, we conducted some evaluations for different components of the proposed *significance* measuring system. One the most important components taken into consideration in this phase was the *significance sentence filter*. The sentence filtering system was tested under two different situations: offline evaluation and live evaluation. For the offline evaluation, the system was tested using 10-fold cross validation. More than 88% of the instances were classified correctly with RMSE of 0.3464. The system was able to classify nearly perfect

Table 4. Classification Results Over Train Data (different features removed)

Parameter assessed	L	S	Fr
Correctly Classified Instances	81%	75%	82%
Incorrectly Classified Instances	19%	25%	18%
Mean absolute error	0.19	0.25	0.18
Root mean squared error	0.4359	0.5	0.4243

over its training data. This is not a complete proof of concept since our data was limited; therefore, the system was also tested under a live evaluation mode in order to determine its generalization capabilities. To test the system, 100 random sentences were selected from the web to test the system's generalization performance. These sentences were hand-labeled and over 72% of the sentences were classified correctly. It can be seen that the results obtained in the live evaluation mode is not as accurate as those of the offline evaluation mode.

Moreover, it is also important to prove that the existence of each feature is important in the results obtained. In other words, features should have positive effects toward the results obtained, and redundant features should be removed. In order to analyze this, each feature was removed once and the system was tested once again over its training data. The results are available in Table 4. In this table L, S, and FR represent the length, signs-of-significance, and frequent patterns, respectively. The removal of signs-of-significance had the most effect whereas the frequent patterns had the minimum effect on our results.

Finally, we evaluated the overall significance measurement of our system. The overall significance measurement is evaluated using a similar method used in topic discovery evaluation. A total number of 665 pairs of posts were used in this evaluation. For each post in the pair, the system gives a significance value which can be used to distinguish which post is of higher significance. The human subjects also chose the post that they believe is of higher significance in each pair. The results showed that the system significance measurement was correct in 58.10% of the pairs. Although this value is not very high, it is considerably above the baseline.

8 Final System Analysis

In this section an overview of the final analysis of the proposed system is provided. The system demonstrates the temporal variation in frequency and significance of different topics. In order to perform this task better, a detailed case study of a social issue and its effect on our system is provided. We have also shown a comparison of the results provided by our system with another famous system in this area, BlogPulse [13].

As described previously, in order to detect the topic of discussion in the blogs, the set of blog posts are clustered and for each cluster a set of keywords are chosen. The chosen keywords represent the topic of each cluster.

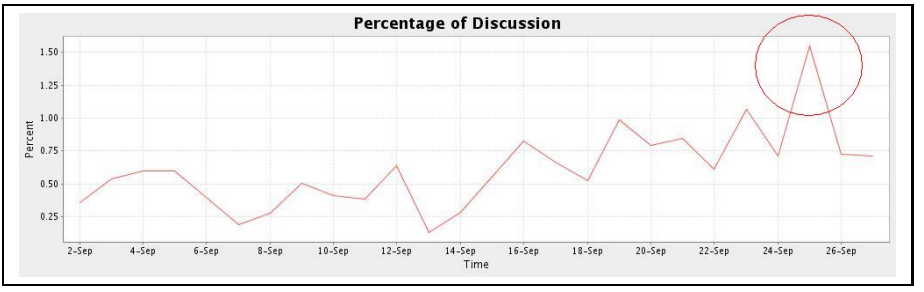

Fig. 4. Frequency Diagram for Topic: *iran*, *iranian*, ...

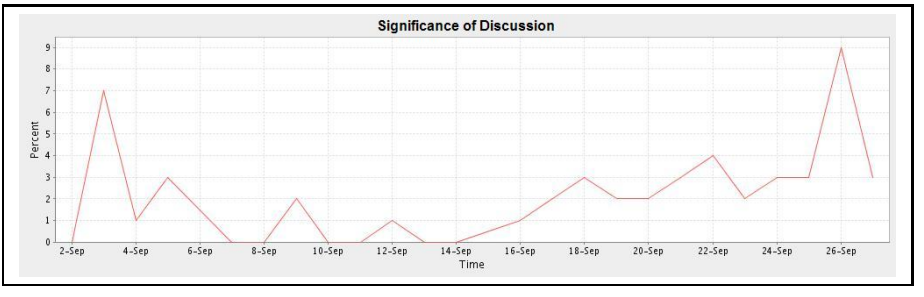

Fig. 5. Significance Diagram for Topic: *iran*, *iranian*, ...

For each topic the frequency of posts are calculated over time. The frequency percentage of a topic is visualized by a diagram, which shows relative amount of discussion on that topic compared to all other discussions in the blogs over time. The frequency diagram for topic (*"iran, iranian, ahmadinejad, jewish, columbia, holocaust, israel"*) is shown in Figure 4.

In addition, for each topic the average significance value is calculated for all posts in that topic and a diagram of its variation is shown to the user. The temporal variations of significance for the same topic is shown in Figure 5.

8.1 Case Study

In order to conduct this case study, we have selected one of the popular social topics found by our system among the blogs under analysis. This topic is identified by these seven keywords: "*israel, iran, iranian, ahmadinejad, jewish, columbia, holocaust*". The significance and frequency diagrams for this topic are taken into consideration and sudden spikes in these diagrams are analyzed. The frequency and significance diagrams of this topic is shown in Figures 4 and 5, respectively. It can be observed in these figures that on the September 24th, 2007, there is a noticeable spike in both textual frequency and significance diagrams of this topic. This spike is due to the Iranian president speech at Columbia

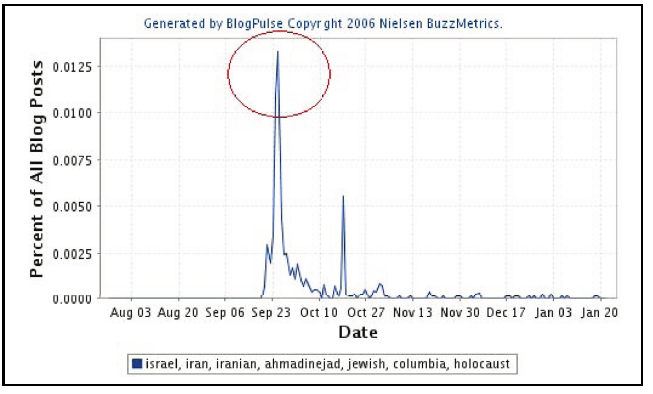

Fig. 6. BlogPulse Trend Analysis for Keywords: israel, iran, iranian,...

University [14]. The same spike can also be seen in similar systems such as BlogPulse. Figure 6 shows the BlogPulse output for the set of keywords available in this topic. The red circles in both Figures 4 and 6 represent the spike date.

It can be seen in Figure 4 that the frequency of discussion is gradually increasing from September 2nd to September 25th. This shows that during this period, the topic is becoming a popular one among authors.

9 Conclusions

In this paper, a novel approach of behavior analysis in blogs has been presented. This approach harnesses the power of significance and frequency measures. Besides these factors, social network analysis has been employed to improve the results inferred from our system. Our detailed analysis and our proof of concept case study provides some promising results. Based on the obtained results, more than 70.52% of our topic assignments and 58.10% of our significance assignments were ascribed successfully.

For the future, we plan to measure the significance and frequency of textual data more accurately and by employing more related features. We also plan to automate the complete process of socio-political issue discovery by means of these features and other possible factors.

Acknowledgments

This work was funded by the Atlantic Canada Opportunity Agency (ACOA) through the Atlantic Innovation Fund (AIF) and through grant RGPIN 227441-00 from the National Science and Engineering Research Council of Canada (NSERC) to Dr. Ali A. Ghorbani.

References

1. Macdonald, C., Ounis, I.: The trec blogs06 collection: Creating and analysing a blog test collection. Department of Computer Science, University of Glasgow Tech Report TR-2006-224 (2006)
2. Dunning, T.: Statistical Identification of Language. Computing Research Laboratory, New Mexico State University (1994)
3. Fellbaum, C.: Wordnet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
4. Porter, M.F.: An algorithm for suffix stripping. Readings in information retrieval, 313–316 (1997)
5. Pelleg, D., Moore, A.: X-means: Extending K-means with efficient estimation of the number of clusters. In: Proceedings of the 17th International Conf. on Machine Learning, pp. 727–734 (2000)
6. Steinbach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques. In: KDD Workshop on Text Mining, vol. 34, pp. 35–36 (2000)
7. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems 30(1-7), 107–117 (1998)
8. Gong, Y., Liu, X.: Generic text summarization using relevance measure and latent semantic analysis. In: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 19–25 (2001)
9. Jindal, N., Liu, B.: Mining comparative sentences and relations. In: AAAI 2006 (2006)
10. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. ACM SIGMOD Record 22(2), 207–216 (1993)
11. Wiebe, J., Breck, E., Buckley, C., Cardie, C., Davis, P., Fraser, B., Litman, D., Pierce, D., Riloff, E., Wilson, T., Day, D., Maybury, M.: Recognizing and organizing opinions expressed in the world press. In: Working Notes-New Directions in Question Answering (AAAI Spring Symposium Series) (2003)
12. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the ACL, pp. 271–278 (2004)
13. Glance, N., Hurst, M., Tomokiyo, T.: BlogPulse: Automated Trend Discovery for Weblogs. In: WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics (2004)
14. Wire, C.T.: President Ahmadinejad Delivers Remarks at Columbia University, Washington Post, September 24 (2007)