Aspect Summarization from Blogsphere for Social Study

Chia-Hui Chang and Kun-Chang Tsai
Dept. of Computer Science and Information Engineering
National Central University, Taiwan
chia@csie.ncu.edu.tw, turninto@gmail.com

Abstract

In this paper, we study the problem of summarizing reasons from blogsphere for social study. We regard weblogs as a source for collecting non-discrete public opinions, where genuine reasons/aspects can be found. To extract the reason inside the blogs, we define four tasks: irrelevant blog filtering, reason/non-reason classification, polarity identification, and reason summarization. We solve the reason/non-reason classification problem by selecting a set of topic related words and brief the reasons by clustering paragraphs containing aspects after sentiment classification. Initial experiments on two topics show an encouraging result on the proposed framework.

Keywords: weblogs, social study, opinion extraction, reason summarization, text mining, sentiment classicization

1. Introduction

Blogs, short for Weblogs, are a frequently updated Internet journal that has become a growing Internet subculture. Blogs have received much attention, because they provide users a place to post their thoughts easily and thus become a mass movement. Nowadays, Weblog becomes the popular place where people can use to express their opinions. According to the report from blog search engine, Technorati, as the end of July 2006, about 1.6 million new weblogs were created each day, which means that there are 18.6 blogs created each second of each day.

The popularity of weblogs opened up a lot of new opportunities in social study. While in the past, researchers in social study had to spent huge efforts in collecting data, they nowadays harvest the data free from the WWW, especially on blogs. For example, we can collect public opinions non-intrusively from the blogosphere by querying articles and comments on a specific topic during a specific time period. However, the large volume of returned documents requires analysis and summarization to help us understand public opinions.

Previous studies on the opinions extraction focus on polarity identification, i.e., sentiment orientation, which distinguish whether authors like or dislike something. For example, Ohno et al. [7], proposed a social summarization method for online auction feedback comments, like book reviews, or hotel reviews. On the other hand, some researches focus on extracting features from product reviews, like movies and hotels. For example, Hu and Liu [3] proposed the idea of extracting what features the reviewers like or dislike.

In the context of social study using online weblogs, where we care about how people think about a topic, knowing whether the blogger approve/disapprove on the topic may not be enough, the reasons why they agree/disagree on it are also important. For example, a sentence like “Some of these babies may have been deformed, have genetic abnormalities, or just be cognitively deficient.” is a positive reason based on eugentics. Such statements are more convincing than “I think it is right to make it illegal.” and “I agree with abortion.”

In this paper, we try to make a summary of a topic by extracting the reasons and aspects behind the general dichotomy that shows people approve or disapprove. There are four tasks for blog analysis: irrelevant blog filtering, reason extraction, polarity identification, and reason clustering. In order to be topic-independent, we proposed an unsupervised approach which selects topic-related words from given blogs and uses the number of topic-related words in a paragraph as an indicator for the strength of containing a reason. Each reason is then judged by the system for its sentiment orientation based on positive and negative words from general inquirer lexicon in the paragraph [17]. In the final stage, positive and negative reasons are clustered separately using frequent itemset based hierarchical clustering, FIHC [2]. This would provide a summary of common reasons why people agree/disagree on a topic. The preliminary result on two topics shows an encouraging performance using the proposed unsupervised approach.

2. Related Work

Opinion summarization on customer reviews, like product reviews [11][1][3], movie reviews [13] has become a hot topic recently. The trend has also extended to editorial reviews [12][4] and blogs [5] as well. The basic problem is a dichotomy of sentiment orientation, either positive or negative. Advanced task is summarizing the texts from different aspects. For example, various features of products/movie like [6][13].

In practice, extracting opinions from customer reviews is easier than from editorial reviews and blogs, because almost each sentence in those articles can be considered as an opinion. In editorial reviews, an additional problem is to distinguish between opinion and fact, i.e., to determine whether a sentence or paragraph is objective or subjective. For blogs, it is even worse because there may be more than one topic in a weblog.

For editorial reviews, Yu and Hatzivassiloglou proposed the use of news as facts, editorials and letters as opinions to train the model to save efforts for labeling training examples. A similar task is from Riloff and Wiebe [9], who separate subjective and objective sentences with supervised machine learning approach. They use some known subjective vocabulary to learn the subjective patterns. The subjective patterns can be used to identify additional subjective sentences, which can augment training data for the extraction pattern learner, which in turn can identify more subjective patterns, and so on.

Sentiment classification has been pursued by two ways. Most of the researches use supervised approach, while others use unsupervised approach [11].

For supervised approaches, the performance varies according to the features. Manually created adjective words with positive and negative labels [12], semantic features based on substitutions and proximity [1], sentiment wordlist and WordNet [4] are the most commonly used resources.

For unsupervised approaches, Turney [11] use internet-based method to get co-occurrence hits of a phrase with “excellent” and “poor” to calculate the PMI-IR scores of the phrase. If the score of the phrase is positive then they consider the phrase is positive, otherwise, the phrase is negative. He then used these phrases to automatically separate positive and negative movie and product reviews, with accuracy of 66–84%.

3. Task Definitions and Algorithms

Figure 1 shows the flowchart of the blog analysis system. The three main tasks are reason extraction, sentiment classification, and reason clustering. Other modules include preprocessing step and irrelevant blog filtering. Given a topic *q* (e.g. abortion), we use google’s blogsearch engine (http://blogsearch.google.com/) to query the first 60 blog entries.

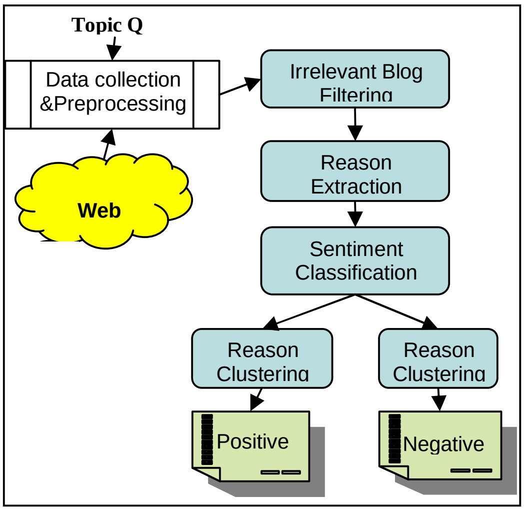

Figure 1: System overview

The pre-processing step contains paragraph segmentation, sentence segmentation, part-of-speech tagging, stemming and paragraph segmentation. We use tags such as <div>, <h>, <p> and consecutive <br> as the rules for segmenting documents into paragraphs, since they are the common tags used by blog service providers. If somehow the blogger didn’t divide his article into paragraphs, then we will not be able to divide it. In practice, almost all bloggers use paragraphs to organize their thought and usually a paragraph contains less than 10 sentences. For sentence segmentation, we use the tool from [13]; to carry out stemming, we use the famous Porter stemmer [15]; for part-of-speech tagging we use GeniaTagger [16].

3.1 Filtering irrelevanat blogs

Although blogsphere provides a non-intrusive method for collecting public opinions, the following problem is the diversified, miscellaneous and complicated opinions which touch many aspects of the query topic, which are way out from the reasons that we want to extract. For example, of the returned blogs for the “gene cloning” topic, some report the pass of bills for using cloning for stem cell research, some are personal profiles but containing the opinion toward cloning without reasons. The mixed contents with irrelevant blogs make the extraction of reasons difficult for our specific study.

To filter such irrelevant blogs, we adopt density based approach which calculates the percentage of sentences containing at least one topic word. If the density is below a threshold, we simply discard that blog.

3.2 Reason extraction

Comparing to feature extraction from product reviews, recognizing the reasons why the authors approve or disapprove on the topic is much difficult. For product reviews, features are generally noun phrases. However, a reason can be expressed in many different ways. Sometimes, a reason can span across sentences to explain. That is why we use paragraphs as the unit for recognizing reasons.

In some way, reason extraction is more similar to the subjective/objective classification problem for editorial reviews. However, we do not have such labeled news and editorials as training examples. Fortunately, we observe that, when people write their reasons, topic words are usually included in the same paragraph. Although, topic words themselves maybe few, they can still be the clues for us to find other topic related words. The density of topic related words in a paragraph can then be used as a measure of how strong a paragraph contains a reason. In this paper, we use logarithm of odds ratio (LODR) score to find topic related words. This measure is then used as an indicator for how closely the word is to the topic we are interested.

To evaluate how related a word *t* is to the topic words, we calculate the probability that *t* occurs in the circumstance that topic words show up as well as the same event when the topic words disappear. We then take the logarithm of the odds ratio of these two probabilities for our first measure. Formally, Let *p* be the probability of the word, *t*, co-occurs with any of the topic words. That is,

$$\begin{aligned} p &= \Pr(\text{ t } \mid \text{paragraph containing topic words}) \\ &= \frac{(\text{\# of paragraphs containing t and any topic word})}{(\text{\# of paragraphs containing any topic word})} \end{aligned}$$

Also, let *q* be the probability of *t*, co-occurs with no topic words. That is, the occurrence probability of the word *t* when no topic word shows.

$$\begin{aligned} q &= \Pr(\text{ t } \mid \text{paragraph not containing any topic word}) \\ &= \frac{(\text{\# of paragraphs containing t and no topic word})}{(\text{\# of paragraphs not containing any topic word})} \end{aligned}$$

The logarithm of the odds ratio is then defined as:

$$LODR(t) = \log \frac{p/(1-p)}{q/(1-q)} = \log \frac{p(1-q)}{q(1-p)} \qquad (1)$$

which is also the difference of the logits of the two probabilities *p* and *q*.

With the weight for a word, we can calculate the average strength of a paragraph containing a reason toward the topic by:

$$ATS(P) = \sum_{t \in P} LODR(t) * f(t \mid P) \qquad (2)$$

where *f(t|P)* denotes the frequency of word *t* in *P*. We use the above average topic strength as an indicator of whether a paragraph contains a reason or not. Generally, the higher the strength, the stronger the paragraph contains a reason.

In additional to the choice of the measures proposed discussed above, another possibility is to focus only on a special set of words like nouns, verbs and adjectives. By excluding other words not in the focused set, we pay more attention to those words with special POS tags. The result will be discussed in the experiment section.

3.3 Sentiment classification

Sentiment classification, or polarity identification, has been the main research problem in opinion extraction. Determining the polarity of a paragraph is more difficult than that for a single sentence because there are more ways to express positions. In this paper, we combine the lexicons used in General Inquirer [17] with Turney's internet-based approach [11] for sentiment classification.

The lexicons in General Inquirer are compiled by experts, with each word labeled with tags, like positive, negative, hostile, strong, etc. A total of 11788 words are included in the vocabulary, while only 4206 words are labeled with positive or negative tags. To extend the dichotomy of words to more grade level, we use Turney's pointwise mutual information (PMI) as the similarity measure of two words. The semantic orientation of a word is then given by the difference between the PMI of a word to "excellent" and "poor".

Let hits(*query*) be the number of hits returned by search engine, such as AltaVista or Google. The semantic orientation of a word can be calculated as follows:

$$SO(phrase) = \log_2 \left[\frac{\text{hits(phrase NEAR excellent) hits(poor)}}{\text{hits(phrase NEAR poor) hits(excellent)}} \right]$$

The word "excellent" and "poor" are chosen for Turney's study because they are mentioned often in product reviews. With the scores for all words in sentiment dictionary, we sum up the scores of the words in the paragraph which contain a reason. If the score is greater and equal than 0, the reason is positive, otherwise, the reason is negative.

$$ASO(P) = \sum_{t \in P \cap D} SO(t) * f(t \mid P) / \mid P \cap D \mid \qquad (3)$$

3.4 Reason Clustering

The purpose of clustering is to provide a good visualization for the reasons extracted from blogs toward a topic. We use frequent itemset based hierarchical clustering, FIHC [2], to cluster those paragraphs containing reasons. Each paragraph is represented by bag of words. We use either topic related words or all words as the bag. This comparison would be used to verify the effectiveness of topic related words identified.

4. Experiments

In experiments, we use the search result blogs as testing data. First, we input a topic that we are interested in, and then we use the topic as query to retrieve the first 60 search result blogs from Google's blog search engine and limit the search results only from spaces.live.com. We only choose two topics to evaluate, "Abortion" and "Same-Sex Marriage", because to prepare the answers is extremely time consuming. For these two topics, we repost those blogs for each topic on the web, make advertisements and invite people to help us label the reasons and modify the answers.

Our experiments can be divided into four parts, filtering irrelevant blogs, reason extraction, sentiment classification and finally, reason clustering. Since the first three tasks are binary classification problems which are determined here by their function values (e.g. topic word density, average topic strength and average sentiment orientation). Therefore, we use either precision-recall curve or ROC curve for the first 3 tasks. For reason clustering, we use accuracy and conformity to measure the performance. For social study, we shall then summarize the main reasons extracted from blogsphere.

4.1 Binary Classification Evaluation

The first part is about irrelevant blog filtering. Topic word density, although is simple provides a satisfied result for irrelevant blog filtering. The best F1 measure is 0.87 and 0.70 for "abortion" and "same-sex marriage" respectively. The ROC curve (true positive rate vs. false positive rate) is shown in Figure 1 with different threshold.

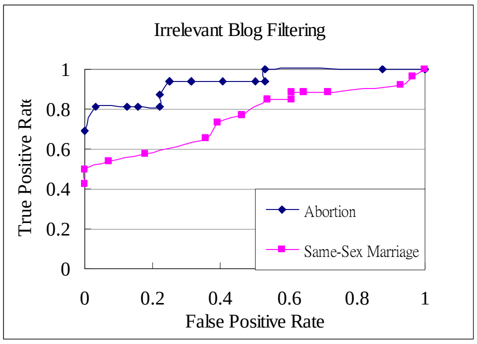

Figure 2: Performance for irrelevant blog filtering

The second part is reason extraction. In this part, each paragraph in the filtered blogs is considered an input. If the average topic strength is greater than a threshold, our system will identify it as a reason. After filtering out the blogs with no reasons, there are still 694 paragraphs and only 88 paragraphs contain reasons in topic "abortion". For "same-sex marriage", there are a total of 701 paragraphs after filtering and only 86 paragraphs contain reasons.

We show precision-recall curve for reason extraction in Figure 3. Performance of topic "abortion" is better than "same-sex marriage", which is consistent with the previous step. The F1-measures for reason extraction are 51.2% and 47.9% for "abortion" and "same-sex marriage", respectively. If we can improve the performance of no reason blogs filtering, performance of reason extraction can be better improved.

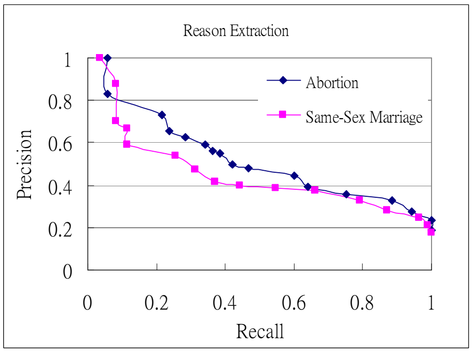

Figure 3: Performance for reason extraction

In the third part, we classify paragraphs containing reasons into positive or negative. We use all the paragraphs with reasons as the input data set for sentiment classification. Paragraphs which contain both positive and negative reasons are ignored in the experiment. Figure 4 shows the ROC curve for sentiment classification. It shows that "same-sex marriage" has better performance than "abortion". One explanation for this is that when people mentioned about the former, the words used are more rational compared to those words used for the later. Particularly, more counter examples are used in the "abortion" topic, which makes the prediction outcome worse.

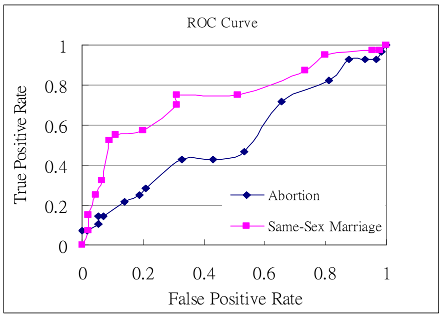

Figure 4: ROC curve for sentiment classification

4.2 Clustering Evaluation

For summarization of the top reasons, negative and positive reasons are clustered separately. We have manually identified reason groups with more than 1 paragraph as main reasons and compared them to reason clustering. Theoretically, reason clustering could group paragraphs with the same reason in one cluster. Meanwhile, information visualization should also present the main reasons to users such that main reasons are identified as early as possible. Therefore, both gold standard (manually) and clustering result (system) are sorted by cluster size and compared via confusion matrix.

Table 2: Confusion Matrix
(the numbers in parentheses denote the cluster size)

	C1 (5)	C2 (4)	C3 (4)	C4 (3)	C5 (2)	C6 (2)	C7 (1)
R1(7)	0	0	**3**	1	0	2	1
R2(4)	0	**3**	1	0	0	0	0
R3(4)	**3**	0	0	1	0	0	0
R4(2)	2	0	0	0	0	0	0
R5(2)	0	0	0	0	**2**	0	0

To evaluate a clustering, we propose two measures: conformity and user view. Given a confusion matrix with *m* real clusters (with size > 1) and *n* prediction clusters like Table 2, where both dimensions are sorted by cluster size, we define conformity as the average largest proportion of paragraphs with the same reason within each prediction cluster (equation 4). We only average over prediction clusters with size greater than 1, otherwise conformity can be maximized when every prediction cluster has size 1. As an illustration, the conformity for Table 2 is $(\frac{3}{5}+\frac{3}{4}+\frac{3}{4}+\frac{1}{3}+\frac{2}{2}+\frac{2}{2})/6$.

$$Conf = \sum_{i=1}^{n'} \frac{\max_j |C_i \cap P_j|}{|C_i|} \times \frac{1}{n'} \qquad (4)$$

There are two kinds of user view: sorted and unsorted. The user view at position *j* is defined as the number of prediction clusters seen by user before R_j (for sorted) or *j* (for unsorted) main reasons are displayed. In other words, the sorted user view at position *j* is the least *s* such that $\bigcup_{i=1}^{s} C_i$ contains at least one reason from each R_i for 1≤i≤j. Similarly, the unsorted user view at position *j* is the least *t* such that $\bigcup_{i=1}^{t} C_i$ contains at least one reason from *j* main reason clusters. As an example, the user view for Table 2 at position 1, 2, 3, 4 and 5 are:

−3, 3, 3, 3, and 5 for sorted user view;
−1, 1, 2, 3 and 5 for unsorted user view;
−1, 3, 4, 5, and 7 for complete unsorted user view;
−7, 7, 7, 7, and 7 for complete sorted user view.

Although the conformity formula excludes cluster with size 1, this measure still favor clustering with many small groups. For user views, this could sometimes affect the value. Generally speaking, the smaller the user views, the better the performance it achieves.

Table 3 shows the performance of reason clustering on positive and negative reasons for both topics. We use either all words or only top *3/4* topic related words, i.e. the words with largest *3/4* LODR score. We also show the sorted userview with respect to different positions in Figure 5. The average number of clusters for both clustering are 18.3 and 19.8 for all words and top 3/4 words, respectively. Although using all words produce a better conformity, it all takes more cluster views to read the major reasons.

Table 3: Clustering Performance Comparison

Conformity		All words	Top 3/4 words
Abortion	Positive	0.559	0.655
	Negative	0.597	0.547
Same-sex Marriage	Positive	0.721	0.661
	Negative	0.619	0.513
Average		0.624	0.594

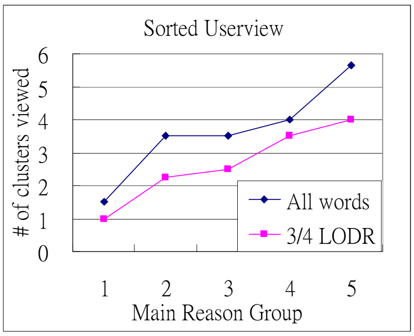

Figure 5: Sorted Userview Curve

4.3 Reason Summarization

In this section, we list the main reasons (with more than 2 supports) from the positive and negative side for both topics. The top positive reasons toward “abortion” are listed as follows. The first reason (with 6 supports) describes that human body is private property and women can decide by herself. The second one (with 6 supports) comes from *eugentics* which concerns babies with serious defect. The third one (with 6 supports) says there might be no support to mothers or they wouldn’t abort. The fourth one explains that some women accidentally got pregnant and abortion should be one of the choices. The fifth main reason claims that some couples believe their family is unbalanced without a son so they want to use abortion.

The first negative reason (with 7 supports) toward "abortion" describes that abortion is not necessary because women can choose adoption instead. The second one (with 4 supports) talks about how cruel the procedure of abortion surgery is. The third one simply indicates that abortion is murder. The forth one explains that there are always alternative solutions to overcome the difficulty. The fifth main reason claims that an unborn baby should be treated as a human with the right been protected.

The top four positive reasons toward "same-sex marriage" are listed follows. The first positive reason (with 10 supports) is that same-sex couple wants to be treated equally. The second (with 4 supports) describe that same-sex marriage doesn't threat heterosexual ones. The third main reason simply states that government has no business in people's private, sexual lives, they can't ban same-sex marriage. The fourth one explains that as long as two people love each other, they can marry no matter what their sexes are.

The first negative reason (with 5 supports) describes that same-sex marriage shouldn't be legal because they often adopt children, and the children in this kind of family will have huge emotional problems. The second reason (with 4 supports) states that church shouldn't be forced to hold same-sex marriage. The third reason illustrates that same-sex marriage is unnatural. The fourth one explains that we should not break the traditional definition of marriage as a union of one man and one woman.

5. Conclusion and Future Work

In this paper, we propose the application of blogsphere on social study for public opinion extraction. Contrast to recent opinion extraction, we stress the importance of reasons for each given topic. We solve the problem of reason summarization by four subtasks: irrelevant blog filtering, reason extraction, sentiment classification and reason clustering. The main contribution of this work is two folds:

–We propose unsupervised approaches for the above subtasks, which make them easy to apply to different topics.

–We suggest two new measures: conformity and user views for clustering evaluation. The greater the conformity and least user views been required for inspection are the better in terms of users' perspective.

The preliminary experiments on two topics show encouraging result for the proposed approaches. Further experiments on more data sets are necessary to prove the effectiveness. Meanwhile, other sophisticated approach can be devised to better improve the performance.

Acknowledgement

This work is sponsored by National Science Council under grant NSC95-2627-E-008-002.

References

1. K. Dave, S. Lawrence, and D.M. Pennock. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. *WWW'03*, pp. 519-528.
2. B. Fung, K. Wang and M. Ester. Hierarchical Document Clustering Using Frequent Itemsets. *SDM'03*, pp. 59-70.
3. M. Hu and B. Liu. Mining and Summarizing Customer Reviews. *SIGKDD'04*, pp. 168-177.
4. S.-M. Kim and H. Eduard. Determining the Sentiment of Opinions. *Coling'04*, pp. 1367-1373.
5. L. W. Ku, Y. T. Liang and H. H. Chen. Opinion extraction, summarization and tracking in news and blog corpora. *AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, pp. 100-107.
6. B. Liu, M. Hu and J. Cheng. Opinion observer: analyzing and comparing opinions on the Web. *WWW'05*, pp. 342-351.
7. H. Ohno, Y. Kusumura, Y. Hijikata and S. Nishida, Social summarization method for feedback comments in online auction. *Syst. Comput. Japan* 37, 8 (Jul. 2006), 38-55.
8. B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *ACL'04*, pages 271-278.
9. E. Riloff and J. Wiebe. Learning extraction patterns for subjective expressions. *Proceedings of the 2003 Conference on EMNLP*, pages 105-112. 2003.
10. H. Takamura, T. Inui and M. Okumura. Extracting Semantic Orientations of Phrases from Dictionary. *NAACL-HLT'07*, pp. 292–299.
11. P. Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *ACL'02*, pp.417-424.
12. H. Yu and V. Hatzivassiloglou. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. *EMNLP'03*, pp. 129–136.
13. L. Zhuang, F. Jing and X.Y. Zhu, Movie review mining and summarization. *CIKM'06*, pp.43-50.
14. Sentence segmentation tool from Cognitive Computation Group in UIUC http://l2r.cs.uiuc.edu/~cogcomp/tools.php
15. http://www.tartarus.org/martin/PorterStemmer/index.html
16. www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/
17. http://www.wjh.harvard.edu/~inquirer/