Diversionary Comments under Political Blog Posts

Jing Wang
University of Illinois at Chicago
jwang69@uic.edu

Clement T. Yu
University of Illinois at Chicago
yu@cs.uic.edu

Philip S. Yu
University of Illinois at Chicago
psyu@uic.edu

Bing Liu
University of Illinois at Chicago
liub@cs.uic.edu

Weiyi Meng
SUNY at Binghamton
meng@cs.binghamton.edu

ABSTRACT

An important issue that has been neglected so far is the identification of diversionary comments. Diversionary comments under political blog posts are defined as comments that deliberately twist the bloggers' intention and divert the topic to another one. The purpose is to distract readers from the original topic and draw attention to a new topic. Given that political blogs have significant impact on the society, we believe it is imperative to identify such comments. We then categorize diversionary comments into 5 types, and propose an effective technique to rank comments in descending order of being diversionary. To the best of our knowledge, the problem of detecting diversionary comments has not been studied so far. Our evaluation on 2,109 comments under 20 different blog posts from Digg.com shows that the proposed method achieves the high mean average precision (MAP) of 92.6%. Sensitivity analysis indicates that the effectiveness of the method is stable under different parameter settings.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval-Information filtering, retrieval models, selection processes

Keywords

diversionary comments, spam, topic model, LDA, coreference resolution, extraction from Wikipedia

1. INTRODUCTION

As a strong force of public opinions, blog comments attract attention from people with different backgrounds. Ideally, commentators write their truthful opinions to help shape and build the contents in the blog posts. However, in practice, various types of unrelated comments are also written by people with different intentions. For instance, companies post advertisements to promote products, and politicians or their supporters leave comments to divert the readers' concerns to another political issue. Many kinds of these unrelated comments in the blogosphere have drawn interests from researchers. One type of unrelated comments has hyperlinks to commercial-oriented pages, and is defined as comment-spam[4]. Various initiatives have been taken to reduce comment-spam. However, we did not find any study on detecting comments that try to deliberately divert readers' attention to another topic. Based on a study of 10,513 comments for 115 political blog posts from Digg.com, we found 39.5% of comments trying to divert the discussion topic. Furthermore, according to the research by Brigham Young University[1], most people who closely follow both political blogs and traditional news media tend to believe that the content in the blogosphere is more trustworthy. Given such significant impact of political blog posts and comments, the existence of the large amount of diversionary comments would have considerably negative effect since they not only twist the blogger's original intention, but also confuse the public. Therefore, we believe it is imperative to identify such comments, especially under the political category.

In this paper, we define comments diverting to unrelated topics as **diversionary comments**, and we focus our work on comments under political blog posts. Based on our observation, there are generally five types of diversionary comments, which are listed below (the type distribution among diversionary comments is also given based on a manually labeled data set of 2,109 comments for 20 randomly chosen blog posts):

Type 1 (63.1%)(Comments diverting to different topics):

Those that twist the post content's intention or purposely divert the discussion to a topic that is different from the content of the post. One example of this type is that given a post which states the risky rush to cut defense spending, commentators write to discuss about reducing social security spending without mentioning defense spending. This action tries to steal the public's attention from defense spending to social security spending.

Type 2 (19.5%)(Comments about personal attack to commentators):

Those that comment on the behavior of some preceding commentators without discussing anything related to the topic of the post. An example of this type is "What's the matter with you? Are you only posting at the very lowest level of threads so you don't deal with responses? "

Type 3 (7.3%)(Comments with little content):

Those that lack content and only contain words such as "lol" and "hahaha". Even though they might express agree-

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

ments or disagreements with the preceding commentators or the content of the blog post, their relatedness to the post content is not clear, and therefore, are considered as diversions.

Type 4 (5.8%)(Comments about the hosting website only):

Those that complain about the poor function of the blog hosting website. We consider them as unrelated to the post content. An example diversionary comment of this type is "Everyone should boycott Digg on Monday". In this comment, "Digg" is the hosting website.

Type 5 (4.3%)(Advertisements):

Those that introduce products or refer to companies, and both of which are unrelated to the post content.

This paper proposes an effective unsupervised method, which aims to rank comments in descending order of being diversionary. The method is based on the intuition that the relatedness between two documents can be measured by their similarity. A related comment should have high similarity with the post content or with the preceding comment it replies to, while a diversionary comment should have low similarities with both the post content and the preceding comment. Our approach tries to first represent each comment and post by a vector, then to use a similarity function to compute the relatedness between each comment and the post, and that between each comment and the comment it replies to, and finally rank comments based on these similarities. However, the following reasons make it a challenging task.

(1) It is difficult to find an accurate representation for each comment and the post. Comments are relatively short and can only offer limited literal information. A simplistic way of applying term frequencies to build document vectors would yield low accuracies, because a related comment may not share enough words with the post, while a diversionary comment may share significant words with the post.

(2) Pronouns and hidden knowledge in the comments and post are other obstacles to accurate representations. Firstly, many commentators use pronouns to represent the person or issue mentioned in the post. Without mapping pronouns to their corresponding proper nouns or phrases, the number of occurrences of the person or issue cannot be captured accurately. Secondly, comments under political posts often indicate political figures or events, which are not explicitly mentioned in the post but are closely related to the post content. Thirdly, many words or phrases, though different, may refer to the same topics. Thus when two comments contain different words but refer to the same topics, their representations are different but ideally should be similar.

(3) A commentator may write to reply to the post directly, but may also write to follow a preceding comment. Most blog hosting websites offer a reply-to hierarchy for commentators. However, many comments do not follow the hierarchy, which makes it difficult to find what a comment replies to.

The main contributions of this paper are as follows:

(1) It proposes the new problem of identifying diversionary comments and makes the first attempt to solve the problem in the political blogosphere.

(2) It introduces several rules to accurately locate what a comment replies to.

(3) It proposes an effective approach to identify diversionary comments, which first applies coreference resolution [3] to replace pronouns with corresponding proper nouns or phrases, extracts related information from Wikipedia[10] for proper nouns in comments and the post, utilizes the topic modeling method LDA [6] to group related terms into the same topics, and represent comments and the post by their topic distributions, and then rank comments in descending order of being diversionary.

(4) A data set, which consists of 2,109 comments under 20 different political blog posts from Digg.com, was annotated by 5 annotators with substantial agreement. Experiments based on the data set are performed to verify the effectiveness of the proposed approach versus various baseline methods. The proposed method achieves 92.6% in mean average precision (MAP)[2]. In addition, its effectiveness remains high under different parameter settings.

2. RELATED WORK

By analyzing different types of diversionary comments, we realize that types 2, 4 and 5 belong to the traditional spam in different contexts. Therefore, we investigate related work on various types of spam detection in this section.

The most investigated types of spam are the web spam [7, 8] and email spam [5, 9]. Web spam can be classified into content spam and link spam. Content spam involves adding irrelevant words in pages to fool search engines. In the environment of our study, the commentators do not add irrelevant words as they want to keep their comments readable. Link spam is the spam of hyperlinks, and comment spam [4, 14] is a form of it, but as we discussed in the previous section, diversionary comments seldom contain hyperlinks. Email spam targets individual users with direct mail messages, and are usually sent as unsolicited and nearly identical commercial advertisements. However, diversionary comments are mostly not commercial oriented and may not contain the same kind of features. In addition, comments are written with a context of the post and preceding comments, while emails are written independently.

Another related research is opinion spam detection[11], though it is not conducted in the blogosphere. Jindal and Liu regard untruthful or fake reviews aiming at promoting or demoting products as opinion spam. They detected spam reviews based on supervised learning and manually labeled examples. They detected untruthful reviews by using duplicate and near-duplicate reviews as the training data. However, diversionary comments are different because they are not necessarily untruthful or fake. In addition, we aim to automatically identify all types of diversionary comments without incurring the expensive task of manually collecting and labeling of training data.

3. COMMENTS DATA ANALYSIS

A standard hierarchy of post-comments in Digg is as follows. Under each post, each comment consists of 4 features (username, written time, comment level, comment content). Comments with "comment level" of $(n+1)$ are designed to reply to preceding comments of level n. In addition, if a comment's level is 0, then it is supposed to reply to the post content directly.

Under such a hierarchy, we believe that a relevant comment can be one related to the post content directly, and can also have a close relation with the preceding comment that it replies to, while a diversionary comment is unrelated to

both the post content and the comment it replies to. Therefore, finding what a comment replies to is necessary for the identification of diversionary comments.

3.1 Finding what a comment replies to

In most cases, a comment at level 0 replies to the post content and a comment at level $(n+1)$ replies to a comment at level n. However, in practice, not all commentators follow such rules when writing comments. Therefore, besides the feature of "level", we need to combine other features such as written time and username to locate a comment's reply-to comment. We use the following heuristics to find a comment's potential reply-to comments.

Assume comment A is at level n and written at time t, while its reply-to comment is written at time t'.

(1) If comment A's content contains the information about username such as "@username$j$", then among comments which precede comment A and are written by "usernamej", the reply-to comment of A is the one that has the smallest positive value of $(t - t')$;

(2) Among all comments which precede comment A and have the level $(n-1)$, the reply-to comment of A may be the one that has the smallest positive value of $(t - t')$;

(3) Among all comments which precede comment A and have the level n, the reply-to comment of A may be the one that has the smallest positive value of $(t - t')$;

(4) Among all comments which precede comment A, the reply-to comment of A may be the one that has the smallest positive value of $(t - t')$, no matter what its level is.

(5) If comment B satisfies condition (1), then B is A's reply-to comment, otherwise, all comments which satisfy any of conditions (2), (3) or (4) are considered as potential reply-to comments. If there is only one potential reply-to comment, we consider it as the final reply-to comment. However, if there are multiple potential reply-to comments, we compare the similarities between the comment and all of its potential reply-to comments. Then among all potential ones, we choose the one that has the largest similarity.

However, some comments reply to the post content directly instead of to other comments. The first comment of the post definitely replies to the post. For each of the other comments which have the level of 0, when its similarity with the post is greater than its similarity with its potential reply-to comments, and is greater than a specified threshold t, we consider it replying to the post directly.

4. DIVERSIONARY COMMENTS IDENTIFICATION

In this section, we present the proposed techniques for identifying diversionary comments. We first explain each strategy that we use to exploit the pronouns and hidden knowledge, and the algorithm we use to rank comments. We then discuss the pipeline of our method.

4.1 Techniques

As we mentioned in the previous section, a diversionary comment is unrelated to both the post content and the reply-to comment. Typical similarity functions such as the Cosine [15] function and the KL-Divergence [12] can be used to measure the relatedness between two documents. However, a simplistic way of utilizing these similarity functions would yield inaccuracies based on the experiment performance, therefore, we add the following techniques to compute the pairwise relatedness more accurately.

4.1.1 Coreference Resolution

Coreference resolution groups all the mentioned entities in a document into equivalence classes so that all the mentions in a class refer to the same entity. By applying coreference resolution, pronouns are mapped into the proper nouns or other noun phrases. If we replace pronouns with their corresponding words or phrases, then the entities become more frequent. For example, a post which talks about President Obama's accomplishments, only mentions "Obama" once, but uses "he" several times. Without coreference resolution, the word "Obama" only occurs once. However, with coreference resolution, "he" will be replaced by "Obama", and the frequency of "Obama" increases. In this paper, we use the Illinois coreference package [3].

4.1.2 Extraction from Wikipedia

When a post talks about President Hu Jintao's visit to U.S., a comment which discusses the foreign policy of China will be considered relevant. However, the post does not mention the word "China", and it does not share any words with the comment. A similarity function such as Cosine which utilizes words in common would yield a small value between the post and the comment. Wikipedia comes to help here, which offers a vast amount of domain-specific world knowledge. In the above example, if we search "President Hu Jintao" in Wikipedia, we will find the information that President Hu Jintao is the current president of the People's Republic of China. However, Wikipedia offers much more knowledge than is needed in the analysis of the post or comments. In order to avoid adding noise, we only pick up anchor texts in the first paragraph from the searched webpage since this information is believed to be most related.

4.1.3 Latent Dirichlet Allocation (LDA) [6]

LDA places different terms, which are related and co-occur frequently, into the same topics with high probabilities. Each term can be represented as a vector of topics. Thus, two related terms which share some topics together will have a positive similarity.

In general, a document-topic distribution can be obtained in the LDA model using Gibbs sampling and it is given by formula (1) [16]:

$$\Theta = \frac{C_{dj}^{DT} + \alpha}{\sum_{k=1}^{T} C_{dk}^{DT} + T\alpha} \quad (1)$$

Here, D and T stand for documents and the number of topics respectively, C_{dj}^{DT} is the number of occurrences of terms in document d, which have been assigned to topic j, and α is a smoothing constant. Based on formula (1), the distribution of a document on a set of topics can be estimated. Then we can compute the similarity between two documents by using topic distributions as vectors.

Using Gibbs sampling, a term-topic distribution is also obtained and it is given by formula (2)[16]:

$$\varphi = \frac{C_{ij}^{WT} + \beta}{\sum_{k=1}^{W} C_{kj}^{WT} + W\beta} \quad (2)$$

Here, W and T stand for the number of terms and topics respectively, C_{ij}^{WT} is the number of times that term i has been assigned to topic j, and β is a smoothing constant.

4.1.4 LDA on training and test data

In order to build an accurate LDA model, a substantial amount of data is required. A post and its associated comments usually have limited amount of data. To obtain enough data, we submit the title of the post as a query to search engines and obtain the first 600 documents as preliminary data to build an LDA model. We denote this data as the training data. Then we build another LDA model on the test data (which is the set of comments of the post), but the term-topic distribution from the LDA model on the training data is utilized in the following way: when running Gibbs sampling to determine the topic assignment for each term occurrence in the test data, if the term has appeared in the training data, its term-topic distribution from the LDA model on the training data is used, but if the term only appears in the test data, the above formula (2) is applied to decide the topic assignment. At the same time, the document-topic distribution for documents in the test data is obtained based on the above formula (1). Then after this LDA model is built, we can use the topic distributions as the document vectors to compute pairwise similarities.

4.1.5 Rank comments in descending order of being diversionary

Algorithm 1 Rank comments in descending order of being diversionary

```
Constants t, t1, t2, t3, t4, where t > 0, t1 ≤ t3, and t2 ≤ t4
for each comment do
  C1 = the similarity between the comment and the post;
  C2 = the similarity between the comment and its reply-to comment;
  if its level = 0 and C1 > C2 and C1 ≥ t then
    C2 = C1;
  end if
  if (C1 < t1 and C2 < t2) then
    Put the comment into potential diversionary list(PDL);
  else if (C1 > t3 or C2 > t4) then
    Put the comment into potential non-diversionary list(PNDL);
  else
    Put the comment into the intermediate list(IL);
  end if
end for
Sort comments in PDL in ascending order of sum(C1, C2);
Sort comments in IL in ascending order of max(C1 − t1, C2 − t2);
Sort comments in PNDL in ascending order of max(C1−t3, C2−t4);
Output comments in PDL followed by comments in IL, followed by comments in PNDL.
```

According to the property that a diversionary comment is unrelated to both the post content and its reply-to comment, if a comment has small similarities with both the post and the reply-to comment, there is a high probability for it to be diversionary. As a consequence, we set two thresholds t_1 and t_2 such that if a comment's similarity with the post (C_1) is less than t_1 and its similarity with the reply-to comment (C_2) is less than t_2, then it is placed into a list called potential diversionary list (PDL). In contrast, if a comment has a big enough similarity either with the post or with its reply-to comment, it is very unlikely to be diversionary. As a result, we set two thresholds t_3 and t_4 such that if the similarity of a comment with the post is higher than t_3, or its similarity with its reply-to comment is higher than t_4, then it is placed into a list called potential non-diversionary list (PNDL). Comments which belong to neither of the above two lists are placed into an intermediate list (IL). Comments in this list do not have high probabilities of being diversionary relative to those in PDL; they do not have high probabilities of being non-diversionary compared to those in PNDL either. Thus, comments in PDL are placed ahead of comments in IL, which are ahead of comments in PNDL.

Based on the above analysis, we use Algorithm 1 to rank comments.

4.2 Pipeline of the Proposed Method

Our proposed method combines all the techniques discussed above to identify diversionary comments. Each step in the procedure is described below:

(1) Submit each post title as a query to search engines and retrieve the first 600 web pages. We extract contents from them as the training corpus. The test corpus consists of each post and the associated comments.

(2) Apply coreference resolution to each document in the training corpus and the test corpus separately, and replace pronouns with their corresponding proper nouns or phrases.

(3) Identify proper nouns in the test data and search them through Wikipedia. For each proper noun, if an unambiguous page is returned, terms in the anchor texts in the first paragraph of the page are added into the document.

(4) Build an LDA model based on the training and test data as discussed in section 4.1.4. The document-topic distribution of each document in the test corpus is obtained.

(5) According to the rules described in Section 3.1, compute the similarity between each comment and the post, and the similarities between each comment and its potential reply-to comments in the test corpus and then decide what a comment replies to.

(6) Rank comments (the test corpus) based on the algorithm in 4.1.5.

5. EVALUATION

For the experiments of this work, we use a data set collected from the politics category in Digg.com, which contains 2,109 comments under 20 different political posts. Each post contains around 100 comments. The corpus is annotated by 5 annotators, and they resolve their disagreement in the annotations together. We consider the final annotation as the gold standard.

5.1 Diversionary Comments Distribution

In this section, we report diversionary comments distribution variation. Based on the gold standard, there are 834 diversionary comment in the test corpus, which account for 39.5% of all comments. We observe that most posts contain 35% to 45% of diversionary comments, and among all diversionary comments, type 1 is the most significant while type 5 takes a relatively low percentage, which also indicates that diversionary comments studied in this work are not commercially oriented, but focus on those deliberately diverting to other topics.

5.2 Experimental Results

As we proposed in section 4, our method consists of several techniques. In order to test the necessity of combining them, we perform experiments by comparing our final method with baseline methods which only apply one technique or combine fewer techniques. The effectiveness of each method is measured by mean average precision(MAP)[13].

In order to keep consistency among all methods to be compared, we set parameters t, t_1, t_2, t_3 and t_4 using fixed percentiles, which are required in the ranking algorithm as

Table 1: MAP for Cosine with different techniques

Techniques /MAP Results(%)	Baseline	With Coref	With Wiki	With Coref Wiki
Term Frequency	69.9	70.4	71.7	72.2
LDA on test data	57.4	58.0	61.1	60.1
LDA on training and test data	75.4	76.7	80.8	92.6

presented in Algorithm 1. In the section below, we set t to 50%, t_1 to 10%, t_2 to 20%, t_3 to 50%, and t_4 to 90%. When LDA is applied, the number of topics is set to 10, α to 0.1, and β to 0.01.

5.2.1 Comparison Results

We compare the following methods firstly: Cosine similarity with term frequency, Cosine similarity with coreference resolution, Cosine similarity with extraction from Wikipedia, and Cosine similarity with both coreference resolution and extraction from Wikipedia. All these methods represent comments and the post by building vectors based on term frequencies. From the first row of Table 1, we observe that Cosine similarity with term frequency has the lowest MAP value (69.9%), while Cosine similarity with both coreference resolution and extraction from Wikipedia performs the best (72.2%). Yet, even the best result is far from being acceptable. The reasons for these poor results are obvious. The Cosine similarity is incapable of matching a document with another document if they have related but different terms. This mismatch can be alleviated to some extent by coreference resolution and extraction from Wikipedia. However, many related terms remain unmatched.

In the second row of Table 1, the LDA model is built simply on the test data, we represent comments and the post by their topic distributions. However, the results are also poor. When coreference resolution, extraction from Wikipedia or both of them are combined with LDA, better results are obtained. However, even the best result has a mean average precision value of 61.1% only. The reason for such a poor result is that the amount of test data is too small for LDA to identify related terms.

In the third row of Table 1, the LDA model is built on the training data and test data, we rank comments based on similarities of their topic distributions. When coreference resolution and extraction from Wikipedia are individually added to LDA, there are notable improvements, but the largest and dramatic improvement comes when LDA and the two techniques are combined, yielding 92.6% mean average precision.

When the Cosine similarity function is replaced by the symmetric KL function, the results turn out be close (89.1%) to those in the third row of Table 1, where the Cosine similarity function is applied.

5.2.2 Sensitivity Analysis

In order to test the stability of our method, we compare its effectiveness by setting different parameters. We first test its sensitivity by setting different numbers of topics while keeping other parameter values unchanged. The number of topics is set to 8, 10 and 12, but similar mean average precisions are obtained. Thus, the method is believed to be stable with different but reasonable number of topics.

Secondly, we test the method's stability by setting different values for ranking algorithm parameters while keeping the number of topics as 10. To make the comparison simple, we set t_1 and t_2 to be the same percentile, and t_4 to be the percentage of t_3 plus 10%. t_1 and t_2 are set in the range from 0.1 and 0.45, while t_3 changes from 0.2 to 0.55, and t_4 changes from 0.3 to 0.65. The average MAP based on Cosine function and symmetric KL are 89.5% and 86.1% respectively. We find that with such wide ranges of threshold values, there is little change in the effectiveness of identifying diversionary comments. Therefore, we conclude that the method is stable with reasonable threshold values.

6. CONCLUSIONS

This paper presented a study on identifying diversionary comments under political posts. In our evaluation data set, 39.5% of comments were annotated as diversions. To the best of our knowledge, this problem has not been researched in the literature. This paper first identified 5 types of diversionary comments, and then introduced rules to determine what a comment replies to under a hierarchy of the post and its associated comments. It then proposed a method to compute the relatedness between a comment and the post content, and the relatedness between a comment and its reply-to comment, which involves coreference resolution, extraction from Wikipedia and topic modeling. We demonstrated the effectiveness of the proposed method using the mean average precision (MAP) measure. Comparisons with baseline methods showed that the proposed method outperformed them considerably.

7. REFERENCES

[1] Brigham young university. http://news.byu.edu/archive09-may-blogs.aspx.

[2] R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. 1999.

[3] E. Bengtson and D. Roth. Understanding the value of features for coreference resolution. In *EMNLP 2008*.

[4] A. Bhattarai, V. Rus, and D. Dasgupta. Characterizing comment spam in the blogosphere through content analysis. *Distribution*, 2009.

[5] E. Blanzieri and A. Bryl. A survey of learning-based techniques of email spam filtering. *Artificial Intelligence Review*, 2009.

[6] D. Blei, A. Y. Ng, and M. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res. 2003*.

[7] C. Castillo and B. D. Davison. Adversarial web search. *Foundations and Trends in Information Retrieval*, 2010.

[8] C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini, and S. Vigna. A reference collection for web spam. *SIGIR Forum*, 2006.

[9] G. V. Cormack. Email spam filtering: A systematic review. *Found. Trends Inf. Retr.*, 2008.

[10] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI 2007*.

[11] N. Jindal and B. Liu. Opinion spam and analysis. In *WSDM 2008*.

[12] S. Kullback. *Information Theory and Statistics*. Wiley 1959.

[13] C. D. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval 2008*.

[14] G. Mishne. Blocking blog spam with language model disagreement. In *AIRWeb 2005*.

[15] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 1975.

[16] M. Steyvers and T. Griffiths. *Probabilistic Topic Models*. Lawrence Erlbaum Associates 2007.