www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 2 February, 2014 Page No. 3982-3985

ENHANCING UNSUPERVISED APPROACH FOR PERSON NAME BIPOLARIZATION

Sathya D[#1], Kaladevi P[*2]

[#]Department of Computer Science and Engineering,
K.S.Rangasamy College of Technology
sathyadurai0301@gmail.com
[*]Department of Computer Science and Engineering,
K.S.Rangasamy College of Technology

Abstract-

A topic is usually associated with a specific time, place, and person(s). Generally, topics that involve bipolar or competing viewpoints are attention getting and are thus reported in a large number of documents. Identifying the association between important persons mentioned in numerous topic documents would help readers comprehend topics more easily. In existing paper propose an unsupervised approach for identifying bipolar person names in a set of topic documents. Specifically, Principal component analysis (PCA) to discover bipolar word usage patterns of person names in the documents, and show that the signs of the entries in the principal eigenvector of PCA partition the person names into bipolar groups spontaneously. To reduce the effect of data sparseness, we introduce two techniques, called the weighted correlation coefficient and off-topic block elimination.

1. Introduction

One benefit is that the web has become an invaluable knowledge base for Internet users to learn about topics. Since the essence of Web 2.0 is knowledge sharing, collaborative tools are designed with the minimum of constraints so that users will be motivated to contribute their knowledge.

As a result, the number of topic documents on the Internet is growing exponentially. To help Internet users comprehend numerous topic documents quickly and easily, topic mining techniques, such as timeline mining, are essential. Existing topic mining approaches focus on extracting important themes in documents of interest. Basically, a topic consists of a sequence of related events associated with a specific time, place, and person(s). Topics that involve bipolar (or competing) viewpoints are often attention getting and generate a large number of documents. However, if people are not familiar with the topics, they may have to expend a great deal of time figuring out the association between important persons mentioned in the documents in order to fully comprehend the topics. Identifying the polarity of the named entities in topic documents, especially person names, would help readers comprehend the topic quickly and easily.

For instance, for American presidential elections, Internet users can find numerous web documents about the Democratic and Republican parties. Identifying the names of important people in the competing parties would help readers form a balanced view of the campaign. In this paper, we define a topic person name bipolarization research method. Given a topic that involves bipolar view points, the method clusters important

persons mentioned in the topic documents into sentiment coherent.

For instance, if theme applied to a set of documents about an American presidential election, it processes the person names mentioned in the documents and identifies important members of the Democratic and Republican parties automatically. Although our research is closely related to sentiment analysis which focuses on discovering bipolar text units mentioned in a set of documents, it differs in a number of respects. Most sentiment analysis approaches identify the polarity of adjectives, adverbs, and verbs. Comparatively few works consider the polarity of named entities. To the best of Specifically, persons with different polarities hold opposite opinions about a certain topic(or issue), while persons in the same polarity group reach a consensus or have the same goal.

Finally, sentiment analysis usually requires external knowledge sources or human composed sentiment lexicons, such as Word Net and General Inquirer, to determine the orientation of a text unit. Sentiment analysis methods normally classify text units in terms of positive orientation or negative orientation, but the polarity of persons may not have positive or negative meanings of a person name is dynamic and context dependent, so no external knowledge source is available for person name bipolarization research. For instance, politicians may agree (or disagree) about a particular topic, but that does not mean they are permanent friends (enemies). The property of context dependence makes the person name bipolarization task a particularly challenging research issue. To resolve the problem, propose an unsupervised approach that identifies bipolar groups of person names in a set of topic documents automatically. Specifically, we use principal component analysis (PCA) to discover bipolar word usage patterns of important person names in a set of topic documents, and show that the signs of the entries in the principal eigenvector of PCA partition the person names in bipolar groups spontaneously. We also present two techniques, called off-topic block elimination and weighted correlation coefficient, to reduce the effect of data sparseness on person name bipolarization. Finally, the occurrences of the identified bipolar person names are organized chronologically to form an active timeline of the topic of interest. As the approach simply analyzes word usage patterns of person names in topic documents, it can be applied to different topic domain sand languages. The results of experiments based on 12 topic document sets written in English and Chinese demonstrate that the proposed PCA-based approach is effective in identifying bipolar groups of person names.

2. RELATED WORK

Our survey of the literature on topic bipolarization revealed that there are surprisingly few related works. This is probably because there search subject this relatively new. Essentially, the technique clusters person names in topic documents in to bipolar groups. In this section, we consider two closely related research subjects, namely, person name clustering and sentiment analysis and also discuss topic time line mining.

2.1Topic Clustering

Topic clustering has attracted a considerable amount of attention inrecent years because using topics to search for information is one of the most popular types of searches on the Internet. However, when are late topic is input to a search engine, the returned web pages may contain information about more than one concept, so it may be difficult for the user to find the desired information. The goal of person name clustering is to facilitate searching with related topics by partitioning the returned web pages into clusters, each of which represents a specific person. The Web People Search (WePS) evaluation work shops provide various data set stop romote the development of efficient and effective person name clustering methods. Most clustering method sare based on the assumption that each returned page refers to a single person. The system uses information extraction techniques to obtain names, job titles, organizations, and e-mail addresses from a

webpage. Specifically, there presentative named entities mentioned in a pair of pages are submitted to a search engine, and the number of returned pages indicates the degree of social similarity of the pages. Songetal. Employed latent semantics analysis techniques to disambiguate person names and observed that name sakes usually have different interests. By comparing the distribution so finterests, modeled by probabilistic latent variables in the web pages, name sakes can be disambiguated. To avoid merging persons with similar interests, the string differences between topic are considered

2.2 SentimentAnalysis

Sentiment analysis, which attempts to identify the polarity (or sentiment) of a word in order to extract positive or negative sentences from documents Hatzivassiloglou and McKeown showed that language conjunctions, such as and, or, but, are effective indicators for judging the polarity of conjoined adjectives. The authors observed that most conjoined adjectives (77.84percent) have the same orientation, while conjunctions that use but generally connect adjectives of different orientations. They proposed a log linear regression model that learns the distributions of conjunction indicators from a training corpus the polarity of conjoined adjectives. Turney and Littman manually selected seven positive and sevennegative words as a polarity lexicon and used pointwise mutual information (PMI) to calculate a word's polarity. A word has a positive orientation if it tends to co-occur with positive words; otherwise, it has a negative orientation.

3. PCA-Based Bipolarization

Principal component analysis is a well-known statisticalmethod that is used primarily to identify the most importantfeature pattern in a high-dimensional data set.

3.1 Sparseness of Text Features

When PCA is used to process textual data, the sparseness oftext features is a major problem. To demonstrate the problem, we collected 411 news documents related to the2009 NBA Finals from Google News 4 and counted how often each person name occurred in the documents and alsoevaluate the NBA topic in the experiment section todetermine if the proposed approach is capable of correctly bipolarizing the person names into the teams that played in the finals. Rank the topics in descending order according to their frequency.

3.2 Weighted Correlation coefficient

The data sparseness problem described in scenario 2 affects many statistical text mining and language models . For person names with the same polarity, data sparseness could lead to underestimation of their correlations because the probability that the names will occur together is reduced. Conversely, for uncorrelated persons or persons with opposite polarities, data sparseness may lead to overestimation of their correlations because they are frequently absent simultaneously from the decomposed blocks. While smoothing approaches, such as Laplace's law (also known as added one smoothing), have been developed to alleviate data sparseness in language models , they are not appropriate for PCA. This is because the correlation coefficient of PCA measures the divergence of person names from their means, so adding one to each person vector entry will not change the divergence. To summarize, data sparseness may influence the correlation coefficient when person names do not cooccur.

4. Conclusion

Topics involving bipolar viewpoints are usually reported by a large number of documents. Thus, identifying bipolar person names in the topic documents should help readers comprehend the topics in a more balanced manner. In this paper, we propose an unsupervised approach that identifies the polarity of person names in topic documents. We show that the signs of the entries in the principal eigenvector of PCA can partition

person names into bipolar groups spontaneously. In addition, we introduce two techniques, namely the weighted correlation coefficient and off-topic block elimination, to address the data sparseness problem. Our experiment results demonstrate that the proposed approach can identify bipolar person names in topic documents correctly without using any external knowledge source. Moreover, the approach is context-oriented, and it can be applied to different languages and diverse topic domains. The results of the present study suggest areas for future research. For example, we observed that some of the evaluated person names possessed neutral orientations. Developing an effective method to identify neutral persons in topics would be worthwhile.

Reference

[1] Chien Chin Chen, Zhong-Yong Chen,(2012) "An Unsupervised Approach for Person Name Bipolarization Using Principal Component Analysis", IEEE Transactions on knowledge and data engineering, Vol. 24, No. 11,pp.1963

[2] Fuchun Peng, James Allan,Ramesh Nallapati, "Event Threading within News Topics", Journal of Emerging Trends in Computing and Information Sciences ,Vol. 2, No. 10,pp.546

[3]Jianfeng Gao , Xiaojun Wan,(2010) , "Person Resolution in Person Search Results: WebHawk" , Journal of Emerging Trends in Computing and Information Sciences ,Vol. 2, No. 8,pp.58

[4]Fabrizio Sebastiani, Zhong-Yong Chen(2009), "SENTIWORDNET : A Publicly Available Lexical Resource for Opinion Mining".

[5] Qiaozhu Mei(2005),"Discovering Evolutionary Theme Patterns from Text An Exploration of Temporal Text Mining", IEEE Transactions on knowledge and data engineering, Vol. 25, No. 10,pp.63

[6] Murthy Ganapathibhotla,(2008) "Mining Opinions in Comparative Sentences", International Conference on Computational Linguistics pages 241–248.