FREQUENCY OF SENTENTIAL CONTEXTS VS. FREQUENCY OF QUERY TERMS IN OPINION RETRIEVAL

Sylvester Olubolu Orimaye, Saadat M. Alhashmi
School of Information Technology, Monash University, Jalan Lagoon Selatan, Bandar Sunway, Malaysia

Siew Eu-Gene
School of Business, Monash University, Jalan Lagoon Selatan, Bandar Sunway, Malaysia

Keywords: Sentential, Frequency, Context, Query terms, Grammar-based, Opinion retrieval.

Abstract: Many opinion retrieval techniques use frequency of query terms as a measurement for detecting documents that contain opinion. However, using frequency of query terms leads to bias in context-dependent opinion retrieval such that all documents containing query terms are retrieved, regardless of contextual relevance to the intent of the human seeking the opinion. This can be described as *non-contextual relevance* problem in opinion retrieval systems such as Google Blogs Search and Technorati Blog Directory. Sentence-level contextual understanding and grammatical dependencies need be considered to ensure documents retrieved contain large proportion of textual contents that have the same underlying meaning with the given query instead of using frequency of individual query terms. Thus, we present specific challenges with state-of-the-art opinion retrieval techniques that rely on frequency of query terms and we propose a grammar-based technique for efficient context-dependent opinion retrieval. We believe our proposed technique can solve the *non-contextual relevance* problem common to opinion retrieval systems, and can be used for context-dependent retrieval such as expert search systems, faceted-opinion retrieval, opinion trend analytic, and personalized web search.

1 INTRODUCTION

Understanding and retrieving human's contextual opinion from subjective contributions (e.g. *relevant* or *non-relevant*) is a complex process, especially when it involves human contributors with diversified styles of making opinionated contributions. By *contextual opinion* we mean, opinions given by some humans closely match the intent of another human seeking such opinions, and by *opinionated contributions* we mean, textual contents expressed as a result of unique human's perception (opinion) about a particular topic. In this paper, the term *opinionated* will be used quite broadly to mean information that contains opinion.

Many opinion retrieval systems avoid computational models that treat opinion as a process of cognitive language understanding (Krahmer, 2010). Particularly, they rely on frequency of query terms or probabilistic measures to detect opinion (Hannah et al, 2007). However, little has been done to show grammatical and contextual understanding of each opinionated contribution (Pang and Lee, 2008), such that frequency of relevant sentence is used to identify opinionated documents instead of frequency of query terms. For example, opinions given in blogs are very dynamic (Liu, 2010), and each blog document may discuss different opinionated topics. Therefore, the use of frequency of query terms to identify opinion without considering the context at which the query terms had occurred may lead to bias in overall opinion score.

Given a particular document, individual query terms may occur at different contexts, yet meet the frequency threshold for a certain opinion target, thus explicitly creates bias in the overall opinion retrieved. We argue that frequency of query terms can not imply subjectivity alone without knowing the context at which query terms must be frequent (Pang and Lee, 2008). For example, these two sentences, "*the fight for academic success*" and "*I*

Olubolu Orimaye S., M. Alhashmi S. and Eu-Gene S..
FREQUENCY OF SENTENTIAL CONTEXTS VS. FREQUENCY OF QUERY TERMS IN OPINION RETRIEVAL.
DOI: 10.5220/0003401206070610
In *Proceedings of the 7th International Conference on Web Information Systems and Technologies* (WEBIST-2011), pages 607-610
ISBN: 978-989-8425-51-5

will fight you to finish", have regular occurrence of the word "fight", which may imply *violence* as an opinion target after a certain frequency threshold, whereas, the word "fight" has appeared in two different contexts respectively. That is, the sentence "the fight for academic success" may imply "passion for academic excellence", and the sentence "I will fight you to finish" may imply "violence".

Thus, we propose a grammar-based approach for sentence-level contextual opinion retrieval using Natural Language Processing (NLP) techniques such as Categorial Combinatory Grammar (CCG) (Baldridge and Kruijff, 2004). For example, CCG analyzes each given sentence to show the underlying grammatical dependencies, and it has a high predictive power for understanding linguistic meaning and interpretation.

2 PROBLEM DISCOURSE

For the purpose of this paper, we are interested in showing specific challenges with many state-of-the-art opinion retrieval techniques that rely on frequency of query terms, and then present an effective grammar-based technique that can understand the context at which opinionated information is needed and to be retrieved. We present different instances whereby the use of frequency of query terms can actually harm the nature of opinionated documents to be retrieved. We argue that opinions are context-dependant (Pang and Lee, 2008) and we believe context-dependent opinion cannot be achieved by using frequency of query terms only.

2.1 Frequency of Terms in Opinion Polarity Detection

Opinion polarity detection techniques have recorded some level of success (Siersdorfer et al, 2010). However, some limitations call for more reliable techniques for effective opinion retrieval. In opinion polarity detection, specific keywords within a document are labelled with a particular polarity (e.g. positive or negative). However, determining an effective way to differentiate between what is positive and what is negative is still an unsolved problem. For example in Sarmento et al (2009), the presence of *ironical* phrases and inverted polarity in opinionated documents led to lower precision for positive opinions with just 77% accuracy. As a result of this, the choice of individual words for polarity detection in opinionated documents is still a big challenge.

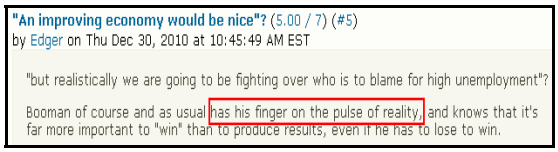

Figure 1: Instance of ironical phrase *"has his finger on the pulse of reality"* in an opinionated document.

2.2 Frequency of Terms in Subjectivity Detection

Subjectivity detection shows if a sentence contains opinion or not (Pang and Lee, 2008). For example, in (Wei and Clement, 2006), Wikipedia knowledge-base was used to identify subjective and objective sentences within a document by considering individual query terms that match Wikipedia concepts. However, this technique encountered *multiple concepts* problem as each query term may return more than one concepts from Wikipedia. We also believe this approach can be computationally expensive as there will be need to iterate through all concepts and articles relevant to each query term. Moreover, the approach assume each query term would always have a single concept, whereas, the given query may sometimes include multi-concepts.

2.3 Frequency of Terms in Lexicon-based Opinion Detection

Lexicon-based approaches consider domain-specific evidences to form lexicons for opinion retrieval (Ding et al, 2008). For generating lexicons, individual opinionated keywords are selected from each sentence in a document. However, we believe opinionated words alone cannot completely and independently express the overall opinion contained in a document, without taking into consideration the grammatical dependencies between words. We argue that the lexicon-based approach is never context-dependent as individual keywords in the lexicon might have been selected from varying grammatical contexts.

2.4 Frequency of Terms in Probabilistic Opinion Detection

Some probabilistic model use general opinion lexicon and proximity density information to calculate probabilistic opinion score for individual query terms (Gerani et al, 2010). We argue that proximity of words to some of the query terms may

not necessarily reflect the context at which opinion is required. In fact, such works assume a single focus opinionated document whereby, opinionated content in a document would explicitly describe the opinion targets without diverging to other possible opinion targets. We argue that this may not usually be true, as most opinionated documents have sentences that express different opinions different opinions even within the same paragraph (Pang and Lee, 2008).

2.5 Frequency of Terms in Language-Model Opinion Detection

The language model combines prediction of occurrence for natural language words and then shows a probabilistic interpretation of such occurrences (Zhai, 2009). For example in Lv and Zhai (2010), proximity heuristic that determine the occurrence of query terms at a distance close to each other was defined, hence determine the document paragraph that has the highest occurrence of such proximity information. Often, this technique requires smoothing procedures and appropriate probability density function. We argue that words in proximity may not determine the context at which human seeks opinion. In fact, it is not yet clear whether such technique can be applied to faceted-opinion retrieval whereby a single document is expected to generate different opinions.

3 EFFECTIVE OPINION RETRIEVAL

Effective opinion retrieval technique can solve the problems identified above. Although, it will be practically challenging to aim at solving all opinion retrieval related problems in a single technique, however, we believe an ideal retrieval system should give sufficient relevance to human's opinionated information need (Pang and Lee, 2008). What makes effective retrieval system is the ability to retrieve opinionated information relevant to the context of the query given, and at a lesser computational cost. Relevant documents retrieved for the opinion target must be relevant to human's intent at a reasonable degree. It should be noted that effective opinion retrieval may not denote a *perfect* retrieval system. For effective contextual relevance purpose, opinion retrieval systems should be able to consider underlying meaning of sentences within opinionated documents. Thus, we argue that a basic retrieval process must aim at reflecting the grammatical context of opinionated information need and not the frequency of query terms.

4 TOWARDS SENTENCE-LEVEL CONTEXTUAL OPINION RETRIEVAL

Towards providing effective solution to the problems identified above, we propose a grammar-based approach for sentence-level contextual opinion retrieval. We believe sentences form the base of the overall opinion being expressed in a document, and opinionated information is better represented in sentences than individual query words. (Pang and Lee, 2008). Therefore, we consider effective opinion retrieval technique that can retrieve sufficient documents relevant to human's opinionated information need. From the series of problems highlighted above, we could observe that the success of any opinion retrieval technique would specifically depend on the degree of relevance to human's intent.

5 GRAMMAR-BASED CONTEXTUAL OPINION RETRIEVAL

We propose to understand the underlying meaning of the given query and each sentence in a given document. For this process, we propose to use CCG which is a *context-sensitive* grammar and NLP technique (Baldridge and Kruijff, 2004). With CCG, we are able to know the underlying meaning and dependencies between words within the given query or the sentences within opinionated documents. The accumulation of sentences that have the same underlying meaning with the given query would determine the contextual relevance of the document to the intent of the human seeking the opinionated documents. By this process, frequency of contextual relevant sentences is used to determine the relevance of opinionated documents instead of frequency of individual query words.

6 LIMITING FACTORS

Knowing the fact that opinions must be detected to match the intent of the human seeking the opinion is only an initiation of idea that must be backed up with practical implementations. However, there are

few limiting factors that may pose major challenges towards the implementation of the context-dependent opinion retrieval approach. By identifying these factors, more research opportunities are created for opinion retrieval. Existing approaches can be improved such that effective opinion retrieval techniques can be achieved in the long-run.

6.1 Dependency Among Opinionated Sentences

Each sentence in an opinionated document may on its own represent certain degree of opinion. However, some sentences depend on prior (i.e. sentence before) or latter (i.e. sentence after) sentences in order to capture adequate opinion being expressed. Detecting context-dependent opinion at sentence-level could as well include looking into dependencies between sentences (Bermingham and Smeaton, 2009). Therefore, opinion retrieval techniques could consider multi-dependency in opinionated sentences.

6.2 Multi-lingual Analysis

Opinionated documents appear in different types of languages (Bruce et al, 2009). Unfortunately, many research works avoid non-English opinionated documents (Kim et al, 2010), while few research works performed bi-lingual analysis of opinions. For example in blogs, contributors have different lingual backgrounds (Bruce et al, 2009), which is why detecting collective opinion without lingual differences is still a general challenge. Therefore, future research works should be aware of this limitation and its significance to the overall opinion retrieval task.

7 SUMMARY & FUTURE WORK

Challenges in state-of-the-art opinion retrieval techniques were reviewed. The cause of major challenges can be summarized as ineffective way of detecting context-dependent opinion at a lesser computational cost. Opinions are context-dependant and a grammar-based opinion retrieval technique is can solve the above mentioned problems. In this paper, we proposed grammar-based approach for detecting context-dependent opinion by using CCG. This approach can be quite useful for faceted-opinion retrieval and personalized web search. In our future work, we plan to implement our sentence-level contextual model for opinion retrieval task.

REFERENCES

Bermingham, A. and Smeaton, A.F. (2009). A study of inter-annotator agreement for opinion retrieval. In *SIGIR'09: Proceedings of the 32nd international conference on Research and development in information retrieval*, pages 784-785. ACM.

Bruce, E. et al., (2009) Mapping the Arabic blogosphere: politics, culture and dissent. *Berkman Center for Internet and Society at Harvard University.*

Ding, X., Liu, B. and Yu, P.S. (2008). A holistic lexicon-based approach to opinion mining. In *Proceedings of the international conference on Web search and web data mining.* Pages 231-240.ACM.

Hannah, D., Macdonald, C., Peng, J., He, B., Ounis, I. (2007). University of Glasgow at TREC 2007: Experiments in Blog and Enterprise Tracks with Terrier. In *TREC*, 2007.

Baldridge , J.M., Kruijiff, G-J.M. (2004). Course Notes on Combinatory Categorial Grammar.

Kim, J., Li, J-J. and Lee, J-H. (2010). Evaluating multilanguage-comparability of subjectivity analysis systems. In ACL'10: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 595-603.

Krahmer, E. (2010) What Computational Linguists Can Learn from Psychologists. A*ssociation for Computational Linguistics*, 36(2): 285-294.

Liu, B. (2010) Sentiment Analysis and Subjectivity. *Handbook of Natural Language Processing*, Second Edition.

Sarmento, L., Carvalho, P., Silva, J.M., Oliveira, E. (2009) Automatic creation of a reference corpus for political opinion mining in user-generated content. In CIKM '09: *Proceeding of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 29-36. ACM.

Lv, Y. and Zhai, C. (2009). Positional language models for information retrieval. In *SIGIR'09: Proceedings of the 32nd international conference on Research and development in information retrieval*, pages 299-306. ACM.

Pang, B. and Lee, L. (2008). Opinion Mining and Sentiment Analysis. In *Found. Trends Inf. Retr.*,2(1-2): 1-135.

Zhai,C. (Statistical Language Models for Information Retrieval: A Critical Review. In *Found. Trends Inf. Retr.*, 2(3):137–213.

Gerani, S., Carman, M.J. Crestani, F. (2010). Proximity-Based Opinion Retrieval. In *SIGIR*, page 978. ACM.

Siersdorfer, S., Chelaru, S. and Pedro, J.S. (2010). How Useful are Your Comments? Analyzing and Predicting YouTube Comments and Comment Ratings. In *International World Wide Web Conference*, pages 891-900.

Wei, Z. and Clement, Y. (2006). *UIC at TREC 2006 Blog Track*, In TREC, 2006.