ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 4, April 2018

Finding Related Forum Post through Synonym Service

Nusaiba C [1], Jaseela Jasmin T. K [2]

M.Tech Scholar, Dept. of CSE, Cochin College of Engineering and Technology, Valanchery, Kerala, India[1]

Asst. Professor, Dept. of CSE, Cochin College of Engineering and Technology, Valanchery, Kerala, India[2]

ABSTRACT: In contrast to traditional approaches for finding related documents that perform content comparisons across the content of the posts as a whole, this paper consider each post as a set of segments, each written with a different goal in mind. The relatedness between two posts should be based on the similarity of their respective segments that are intended for the same goal, i.e., are conveying the same intention. This means that it is possible for the same terms to weigh differently in the relatedness score depending on the intention of the segment in which they are found. A segmentation method that by monitoring a number of text features can identify the parts of a post where significant jumps occur indicating a point where a segmentation should take place. The generated segments of all the posts are clustered to form intention clusters and then similarities across the posts are calculated through similarities across segments with the same intention. Finally a synonym service is used to check the synonomically related forum posts to the post at hand.

KEYWORDS: Segments, Relatedness Score, Features, Clustering, Synonym Service.

I. INTRODUCTION

Forums in the context of online user communities offer users the ability to seek solutions and make decisions regarding diverse problems by exploiting other users' experience. They also offer businesses the ability to connect and support their customer base. The organization of the forum posts into categories is a feature that helps users to identify more easily those posts related to a topic. However, since browsing a very large number of posts is frustrating and time-consuming, most forum sites offer keyword search capabilities. Yet, keyword search may not result in a complete set of related posts since the selection of the right keywords is not always straightforward. For better support users, an important functionality is to provide them with a number of pertinent posts once they have identified a post of interest.
This paper deal with the problem of finding forum posts related to a post at hand. Relatedness has traditionally been translated into content similarity. Content similarity computed directly across forum posts is, unfortunately, not very effective in this case because searches are done under specific thematic categories, e.g., printers, or hotels in New York, in which the content of all the posts is anyway similar.
This system uses segmentation to measure the relatedness of two forum posts, by treating them as composite objects instead of monolithic entities can lead to more effective comparisons. Indeed, a forum post consists of parts, each serving a different goal, i.e., expressing a different message to the reader through the text. For instance, a part may serve to describe a problem that the author has, another to provide background information in order to put the reader into context, a third to express a desire, and a fourth to reach a conclusion. After the use of segmentation the created segments are grouped so that the accurate relatedness can be calculated. Each cluster are the given particular weights using standard weighing schemes. Finally the weights corresponding to each forum are compared to get more accurate related forums. The TFIDF matching algorithm is used for the matching purpose and further the TFIDF algorithm is extended as a synonym based TFIDF to get more accurate result. Once the related forums are found by using TFIDF matching algorithms, then they are compared with a synonym dictionary service to check whether there is synonym icily related forums.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 4, April 2018

II. RELATED WORK

Short text classification by multigranularity test is a main related work. It deals with dividing the short text into different granularity and the granularity taxonomy is used for the comparison purpose. But it only deals with the short text not a complete forum can't be granularized. Next is finding similar questions based on their answer, here some text mining algorithms like LDA are used for finding the similarity between two answers and the compared with their questions. Most of the related papers only contains on of the popular text mining algorithms like LDA, TFIDF, TRLM ect. But they are alone not sufficient for a large forum comparison and for getting the accurate result needed. Some papers deals with how to suggest questions in an online forum. But it is not necessary to get knowledge of how to suggest a question, because the question is asked by the needed one. This paper is extended version of the TFIDF for better result than using one of the text mining algorithm.

III. PROBLEM DEFENITION

The challenge we propose to address is as follows: given a collection D of documents, and a reference document dq, find those k documents in the collection that are most likely to be related to the reference document dq, i.e., those documents that will most likely be of interest to a user that already considers dq being of interest. The specific task is referred to as document matching.

IV. METHODOLOGY

1. Stop words are eliminated for the better result computing and making the computations easy. Stop words are usually referred to as the most common word in a language; there is no single universal list of stop words. Stop words are generally thought to be a single set of words. It really can mean different things to different applications. In order to perform the segmentation of the posts first the feature selection is performed. Features are keywords which are used for identifying the intentions. For a document d, there are $2(|d|-1)$ possible segmentations.
2. The next step in intention-based post matching is to recognize segments that are intended for the same goal (or purpose). We actually need to create groups such that segments with similar intentions end up in the same group and segments with different intentions in different groups. Since the actual intention is not known but we have modelled it through a vector of features, a natural choice for creating the desired groups is to perform clustering on the feature vectors corresponding to the intentions of the segments. Each cluster can then be seen as a representative of some communication goal.
3. To perform the document matching, i.e., to identify the documents in a collection that are related to a reference document dq, one way is to see the document dq as a query and then measure the relatedness of each other document d0 to that query in a way similar to how IR techniques work. As already mentioned, our position is that such a task should not consider each document as a whole but should be specialized on each intention individually, and then combine the results.
4. Each cluster is the projection of every document on the specific intention that the cluster represents. Thus, to measure the relatedness of a document d0 to the reference document dq with respect to a specific intention I, it is enough to measure the relatedness of the respective segment s0 of d0 in the cluster I, to the respective segment sq of dq in that same cluster. For computing this relatedness any text comparison one of the best-known IR techniques is the TF/IDF. The core of the original TF/IDF method and its probabilistic variance BM25 consists of a term weighting scheme that weighs a term in a document considering the number of its appearances in relationship to the number of its appearances in all the other documents. We devise a version that is somewhere between the original and the BM25, and takes into consideration intentions. In particular, we start with a variance of TF/IDF that comes close to BM25 and has been implemented in SQLyog for full text searching.
5. The top-n lists generated across the different intentions, i.e., the set M mentioned above, are used to generate the k most related documents to the reference document dq. A new list R is created that contains every

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 4, April 2018

document that appears at least in one of the lists in M. A score is associated to each such document that is the sum of the scores with which this document appears in the various lists in M. The k elements in R with the highest score are returned as answer to the request of the matching documents to the reference document dq.

6. Finally Synonyms-Based Term weighting scheme (SBT) to make the keyword extraction more effective by overcoming the limitations of tf.idf weighting scheme using a thesaurus of the respective domain. Synonyms Based Term weighting scheme (SBT) which changes Inverse Document Frequency (IDF) according to the synonyms based cluster of any term. A dictionary of synonym is created and each term in the segments are compared with the synonyms in the dictionary for accurate result.

V. CONCLUSION

We have proposed a novel approach for matching a reference post to the k most related posts in a collection. Our method identifies and exploits post segments that convey similar author intentions. We presented several experiments regarding the right segmentation criteria, the effectiveness of the segmentation algorithms and the formation of intention clusters that prove that a rather intuitive concept, that of the author intentions to communicate a certain message, can be effectively captured by an automated process. Moreover, due to the nature of the posts, measuring the relatedness score after having distinguished the different segments/messages that the authors intend to communicate has been proved more effective than the direct comparison of the whole posts.

REFERENCES

[1] M. Chen, X. Jin, and D. Shen, "Short text classification improved by learning multi-granularity topics," in IJCAI, pp. 1776–1781,2011.
[2] J. Jeon, W. B. Croft, and J. H. Lee, "Finding semantically similar questions based on their answers," in Proceedings of the 28th ACM SIGIR Conference, ser. SIGIR '05. New York, NY, USA: ACM, pp. 617–618,2005.
[3] T. C. Zhou, C.-Y. Lin, I. King, M. R. Lyu, Y.-I. Song, and Y. Cao, "Learning to suggest questions in online forums." in AAAI, 2011.
[4] L. Weng, Z. Li, R. Cai, Y. Zhang, Y. Zhou, L. T. Yang, and L. Zhang, "Query by document via a decomposition-based two-level retrieval approach." Association for Computing Machinery, Inc., July 2011.
[5] V. Govindaraju and K. Ramanathan, "Similar document search and recommendation," Journal of Emerging Technologies in Web Intelligence, vol. 4, no. 1, pp. 84–93, 2012.
[6] S. Robertson, S.Walker, and M. Hancock-Beaulieu, "Okapiat TREC-7: Automatic ad hoc, filtering, VLC and interactive track," TREC '98, pp. 199–210, 1998.
[7] D. M. Blei, "Probabilistic topic models," Commun. ACM, vol. 55, no. 4, pp. 77–84, Apr. 2012.
[8] J. Berant and P. Liang, "Semantic parsing via paraphrasing." in ACL (1), pp. 1415–1425,2014.
[9] H. Wen, W. Zhongyuan, W. Haixun, Z. Kai, and Z. Xiaofang, "Short text understanding through lexical-semantic analysis," in IEEE ICDE, 2015.
[10] Z.-Y. Ming, T.-S. Chua, and G. Cong, "Exploring domainspecific term weight in archived question search," in Proceedings of the 19th ACM CIKM, ser. CIKM '10. New York, NY, USA: ACM, pp. 1605–1608,2010.
[11] H. Wu, W. Wu, M. Zhou, E. Chen, L. Duan, and H.-Y. Shum, "Improving search relevance for short queries in community question answering," WSDM, pp. 43–52, 2014.
[12] H. Wu, W. Wu, M. Zhou, E. Chen and L. Duan, H.-Y,2014, "Improving search relevance for short queries in community question answering",in WSDM, pp. 4352, 2014.