BLOG CLUSTERING ENHANCED BY MINING THE WEB COMMENTS

Tomas KUZAR, Pavol NAVRAT

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
tomas.kuzar@gmail.com, navrat@fiit.stuba.sk

Abstract. Web environment gives us an opportunity to take broader aspects of textual information into account, for example information about behavior of web users, global web trends or social interactions between users. In this paper we present and evaluate method suitable for enhancing blog clustering using the information hidden in web comments. We found out that blog clusters based on clustering of the commentators can differ significantly from the content clusters. But according to the results of our experiments implicit relations between commetators can be used in addition to content clustering and improve the quality of content clusters.

1. Social Web

Web 2.0 opened for every web user the option to become a web publisher. Describing someone as the user – or web user – has never been very accurate, but more and more it becomes too abstract. A person formerly described as (web) user assumes at least two different roles: content producer and content consumer. A content producer typically contributes to web space by writing statuses, comments or tags. A content consumer typically is curious, seeks some information and when finds something relevant, reads it. Another typical feature is that both these roles are usually intertwingled, thus a curious fellow acts both as content producer and content consumer at the same time.

More specifically, on the consumer side, a curious fellow should receive content that is as relevant as possible to her information need. This should be achieved in an effective way. Our contribution to achieving these objectives is based on the hypothesis that clustering the web content has a potential for improvement. Our idea is that on the producer side, annotations or comments can be very productive in increasing quality of clusters of documents.

2. Results

In our research we focus on Slovak blog clustering using information mined from the web. Our blog content clustering method consists of several steps: preprocessing (tokenization, stemming), processing (LDA based dimensionality reduction, K-means clustering). Clustering method based on implicit relations creates clusters of the blogs according to commentators who commented the blogs. This method is based on our assumption that similar commentators comment similar blogs. We also assume that blogs have characteristic themes and characteristic commentators. This method does not rely on language processing.

We focused on using web comments in post-processing phase of blog clustering. We proposed and evaluated a method for post-processing of blog clusters which relies on user's comments. In our experiment we combined content clustering with implicit ties between users based on comments. According to the results of our experiments we can claim that clusters based on implicit ties between commentators significantly differ from the content clusters. But the quality of content clusters can be improved by considering implicit ties between commentators in case of articles which do not fit into single cluster.

3. Conclusions

We proposed to organize social web data into content clusters. This can help the curious fellows (information consumers) effectively receive the information they seek. We devised a content processing framework for building high quality content clusters. Processing framework consists of several components which operate in different processing steps. Intensive processing of unstructured data is a prerequisite of high quality outputs. We found out that information preprocessing significantly influences the quality of content clusters. Then we found out that quality of content clusters can be influenced by dictionaries but the domain of the dictionary needs to match domain of the input dataset. We intensively analyzed web comments. We used content of web comments as extension of web documents and also we used web comments as indicator of relations between social web documents. Enriching web documents with comments had positive impact on overall quality of clustering. However, we found that processing comments increases quality of clusters only in cases that comments are processed together with articles.

Acknowledgements: This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0971/11 (2011-2014) and the Slovak Research and Development Agency under the contract No. APVV-0208-10. This contribution/publication is the partial result of the Research & Development Operational Programme for the project Research of methods for acquisition, analysis and personalized conveying of information and knowledge, ITMS 26240220039, co-funded by the ERDF.

References

to other papers publishing the results that are summarized here

[1] T. Kuzar, P. Navrat. Burst Moment Estimation for Information Propagation. In *Advances in Intelligent Web Mastering - 2: Proceedings of the 6th Atlantic Web Inteligence Conference - AWIC 2009*, Czech Republic, (2010), Springer, pp. 147–154.

[2] T. Kuzar, P. Navrat. Preprocessing of Slovak Blog Articles for Clustering. In *Proceedings of the 2010 IEEE/WIC/ACM Int. Joint Conference on Web Intelligence and Intelligent Agent Technology* (WI-IAT 2010) Workshops Proceedings, , Canada, IEEE CS, (2010), pp. 314–317.

[3] T. Kuzar, P. Navrat. Slovak Blog Clustering Enhanced by Mining the Web Comments. In *Proceedings of the 2011 IEEE/WIC/ACM Int. Conferences on Web Intelligence and Intelligent Agent Technology - Volume 03* (WI-IAT 2011), IEEE CS, (2011), pp. 293–296.

[4] Kuzar, T. Clustering on Social Web. *Information Sciences and Technologies Bulletin of the ACM Slovakia*, Vol. 5, No. 1 (2013) pp. 34-42.

Other references

[5] Trivedi A.: Exploiting tag and word correlations for improved webpage clustering. In: *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, ACM, (2010), pp. 3-12.