Abstract

Traditional clustering algorithms are usually based on the bag-of-words (BOW) approach. A notorious disadvantage of the BOW model is that it ignores the semantic relationship among words. As a result, if two documents use different collections of core words to represent the same topic, they may be assigned to different clusters, even though the core words they use are probably synonyms or semantically associated in other form and other disadvantage of conventional web page clustering technique is often utilized to reveal the functional similarity of web pages. Tagging can be beneficial to improve the clustering performance. Several efforts have been made to explore social tagging for clustering. But there is some drawbacks of tagging web based clustering. To our knowledge, all the existing approaches exploiting tag information for webpage clustering assume that all the WebPages are tagged, which is a somewhat restrictive assumption. In a more realistic setting, one can only expect that the tags
will be available for only a small number of WebPages. In this paper, we propose a new web page grouping approach based on Probabilistic Latent Semantic Analysis (PLSA) model. An iterative algorithm based on maximum likelihood principle is employed to overcome the aforementioned computational shortcoming.

References

- Anusua Trivedi, Piyush Rai, Scott L. DuVall "Exploiting Tag and Word Correlations for Improved Webpage Clustering" SMUC'10, October 30, 2010, Toronto, Ontario, Canada. Copyright 2010 ACM.
- http://www.stumbleupon.com
- http://www.delicious.com
- Open Directory Project (http://www.dmoz.org/)

Index Terms

Computer Science  Emerging Trends in Technology

Keywords