Effective Representation for Content-Based News Recommendation

Dušan ZELENÍK*

Slovak University of Technology
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
dusan.zelenik@gmail.com

Our work is based on advantages which could be achieved by the hierarchical representation of similarity between entities. As long as we are working with news we focused on representing similarities among text documents. Our method for similarity search composes a tree of news articles based on content similarity following the related work made by Sahoo [2]. In contrast to mentioned work, we preserve hierarchy especially for the lowest level of the tree where entities are not clustered explicitly, but considered as single entities. This way is our representation ready to produce clusters of similar entities on every level of cluster density. This representation grows incrementally and produces hierarchy, what is effective for growing datasets and dynamically changing domains, because of its logarithmic complexity of storing and retrieving similar articles [4].

In a connection to tree composing we had to solve problem of deep tree form. This form emerges when articles submitted to hierarchy are hardly similar. We use tree balancing to preserve homogeneity of the tree. Advantages of tree balancing are useful in domains where features describing entities are very rare, simple and intersection of features is small. Image tags or keywords extracted from text are then sufficient to compose a tree.

We apply our solution on the domain of news as a part of SMEFIIT project [1], where is the complexity of the method important. Since news are continuously published on mentioned news portal, articles became significantly time sensitive. Therefore, should be service for similar news providing up-to-date. We keep similarity of articles in the hierarchy, so the set of the most similar articles is retrieved very fast. Furthermore are all features of processed articles considered (including newly added).

To prove that our representation is not only fast by also reliable, we evaluated it in comparison with brute force similarity search. We achieved relatively high precision for top similar articles [3]. Articles with lower, but still considerable similarity are

* Supervisor: Mária Bieliková, Institute of Informatics and Software Engineering

Spring 2010 PeWe Workshop, April 18, 2010, pp. 39–40.

omitted, what is consequence of non-existing transitive relations among extracted feature sets.

Finally, we utilized mentioned representation to generate recommendations for users - readers of the news website. It is important to provide recommendations real-time in such a domain. Otherwise, the user loses his patience very easily. The user demands correct and relevant results but keeps waiting very shortly. Complications occur especially in domains where the subject of recommendation is time sensitive and the dataset grows. Our representation utilized for the news recommending solves these issues, because of its low complexity.

One of the advantages is ability to generate recommendations according to content of articles. Another advantage is that recommendations are personalized. The content of an article is mapped on user's interests, which means that articles similar to the articles interesting for user are recommended. We solved complex problem of processing amount of articles to generate recommendations using our effective representation of similarities.

We use incrementally composed hierarchy of similar articles also as a hierarchy of user's interest stereotypes. Each stereotype is a tree node with set of ancestors - similar articles. Since the user reads specific types of articles, we presume that his interest stereotypes could be located in our representation and ordered by its relevance. The ratio of articles read and articles not displayed is a criterion for such a sorting. In a result, the recommendation consists of articles from more relevant stereotypes to cover all of the user's interests. Recommending such a mixture is better, especially because of multivariate nature of single reader and his interest. The content similarity is then effectively used to recommend newly added articles if relevant for specified interests of reader, even with mentioned drawback of omitting less similar articles.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0508/09.

References

[1] Barla, M., Kompan, M., Suchal, J., Vojtek, P., Zeleník, D., Bieliková, M. News recommendation. In Proc. of the 9th Znalosti, pp. 171-174. 2010.

[2] Sahoo N., Callan J., Krishnan R., Duncan G., and Padman R.. Incremental hierarchical clustering of text documents. In CIKM '06: Proc. of the 15th ACM int. conf. on Information and knowledge management, NY, USA, 2006.

[3] Zeleník, D. News Recommending Based on Similarity Relations. In Student Research Conference 2010. 6th Student Research Conference in Informatics and Information Technologies Bratislava, April, 2010 : Proceedings in Informatics and Information Technologies. STU v Bratislave FIIT, 2009.

[4] Zeleník D. and Bieliková M. Dynamics in hierarchical classification of news. In WIKT '09: Proc. of the 4th Workshop on Intelligent and Knowledge oriented Tech. (WIKT 2009), pages 83–87, Kosice, Slovakia, 2009.