VECTOR-BASED TREE NEWS RECOMMENDATION

Mária BIELIKOVÁ, Michal KOMPAN, Dušan ZELENÍK

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
{bielik,kompan,zelenik}@fiit.stuba.sk

Abstract. The amount of information over the Web is increasing day by day. Thanks to the information accessibility increase, users are not only information consumers anymore, but we are involved into the dynamic content creation process. The typical example of the information overload problem is the domain of news. Users are generally overwhelmed by hundreds of thousands of articles, and cannot access relevant information in sufficient time. The personalized recommendation are used to overcome these problems generally. We propose a novel approach for content-based recommendation, where the article structure is considered in order to perform similarity search and highly efficient hierarchical tree structure for the information store is used.

1. Motivation

There is often problem of information overload in the newspaper domain. Plenty of articles are published daily, while users cannot access relevant information easily. Personalized recommendations are often used to minimize amount of information.

Several approaches for recommendation and filtering have been proposed since early nineties. Two approaches for the personalized recommendations have been proposed. The content-based recommendation tries to identify similar items (based on the content) in order to recommend similar items. The collaborative filtering tries to identify similar users in order to recommend items liked by these similar users. These two basic approaches are often mixed to bring better results [4,5].

As the newspaper domain is characteristic with great number of articles, which have to be processed as quick as possible and plenty of users respectively, some kind of optimization of recommendation approaches have to be performed in order to face up these domain dependent specifics.

2. Vector representation

Content-based recommendation is suitable for well-structured domains [3]. The domain of newspaper thanks to the rich metadata describing each article is typical example of such structured domain. On the other side, the one of the main content-based recommendation shortcomings is the computation cost in the mean of the similarity search – article to article similarity have to be computed. This is characteristic for the newspaper domain as well. Number of articles appearing in every hour is tremendous. Because of this some efficient computation and representation for the articles have to be used.

We propose effective vector-based article representation which is based on the available metadata for each article (Table 1).

Table 1. Proposed vector-based article representation.

Title	TF of title words in the content	Keywords	Category	Names/ Places

The article title refers to the pre-processed words from the article title. Second part of proposed vector is focused to map article title to the article content (based on the term frequency). The keywords part consist of top N keywords generated based on the whole corpus of articles. As the every article is assigned to some category in order to filter most similar article, we include the category as one part of proposed representation. The hierarchy of categories over the news portal can be found, we use this information to assign specific weights to each category (Figure 1). Finally, the relevant names and places are extracted in order to match articles about one person in various categories or topics.

We proposed an extension to proposed vector based recommendation to include not only the single words but the words collocations which improves the similarity search over the articles.

	C1	*C2*	*C3*	*C4*
A	1/4	1/2	1	-
B	1/4	1/2	1	-
C	1/4	1/2	-	1

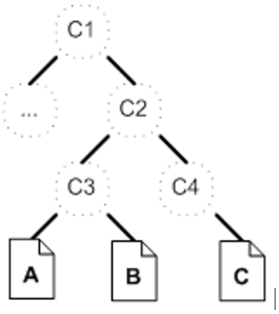

Figure 1. Weights construction for specific categories based on the portal category hierarchy.

3. Tree representation

As the most important feature for the content-based recommendation is fast similarity computation and accessing this information, we propose tree-based representation of similarity of articles. We use binary tree to represent these similarities. The hierarchy itself is generated over these articles and is used for representation of metadata to access similar articles. We designed our representation as a hierarchy where:

- real articles are placed at the lowest level of the tree as leaf nodes,
- features are spread to upper levels of the hierarchy structure,
- similarity is stored in the hierarchy itself.

As the tree is built incrementally the computational cost is reduced significantly, as no more article to article similarity computation have to be computed. Moreover, our tree representation includes special nodes – metadata nodes, which aggregate article metadata from following nodes. Thus after pre-processing article and constructing vector-based recommendation described above we:

- locate a place in the tree where to add the article,
- add the article at the correct place and adjust edges in the hierarchy,
- modify the rest of the structure which is affected by the new article.

The best position search over the three for an article starts in the root. Next step by step comparison in every node (based on the Jaccard similarity, when more similar article wins). Tree traversal ends when a leaf node is reaches or the calculated similarity is nearly equal for each branch. When the best position is found, we append a new node for the new article a spread new meta-information to the parents and root (Figure 2).

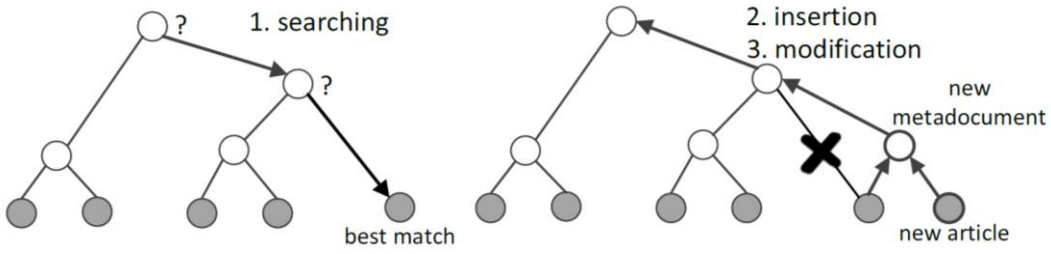

Figure 2. In the first step we find the best place to insert a new article (left picture), which is added at the correct position in the second step. Metadata are created for the new node (right picture) and features are propagated to the root.

4. Conclusions

As the amount of information increases day by day, the personalized recommendation approaches importance increase as well. There are often various limitation, which result to the recommender system failure. In the news domain the amount and the frequency of new articles represent one of such limitation. Here it is crucial to recommend articles as fast as possible, as generally old articles are not so interesting for readers.

We proposed approach for the computation cost reduction for similarity search and recommendation of news articles. Our approach focus to efficient article representation, while vector based representation of specific features ensures the reduction of metadata for every article, while the important information for the similarity search is ensured.

This efficient representation is used in the tree-based storage of "mined" information (article similarities). The binary tree allows us to reduce the number of comparisons, while the dynamical balancing is involved. This allows us to compute article similarities in incremental manner and which is most important in almost real-time.

Proposed approaches bring high computation cost improvements, while the qualitatively improvement can be observed as well (based on the F1 measure).

Acknowledgement: This work was supported by grants No. VEGA1/0508/09, VG1/0675/11, APVV 0208-10 and it is a partial result of the Research and Development Operational Program for the projects Support of Center of Excellence for Smart Technologies, Systems and Services, ITMS 26240120005 and Research and Development. Operational Program for the projects Support of Center of Excellence for Smart Technologies, Systems and Services II, ITMS 26240120029, co-funded by ERDF.

References

to other papers publishing the results that are summarized here

[1] Bieliková, M., Kompan, M., Zeleník, D.: Effective Hierarchical Vector-Based News Representation for Personalized Recommendation. *Computer Science and Information Systems*, (2012), vol. 9, no. 1, pp.303-322.

[2] Kompan, M., Bieliková, M.: News Article Classification Based on a Vector Representation Including Words' Collocations. In: *Proc. Of the 3th International Conference on Software, Services and Semantic Technologies*, Springer, (2011), pp. 1-8.

Other references

[3] Billsus, D., Pazzani, M.: User Modeling for Adaptive News Access. *User Modeling and User-Adapted Interaction*, vol. 10, nos. 2-3, (2000), pp. 147-180.

[4] Bouras, C., Tsogkas, V.: Personalization Mechanism for Delivering News Articles on the User's Desktop. In *Proc. of the 4th int. Conf. on Internet and Web Applications and Services (May 2009)*. ICIW. IEEE Computer Society, Washington, DC, (2009), pp. 157-162.

[5] Melville, P., Mooney, R., Nagarajan, J.: Content-boosted collaborative filtering for improved recommendations. In *Proc. of 18th National Conf. on Artificial intelligence* (Edmonton, Alberta, Canada). AAAI, Menlo Park, CA, (2002), pp. 187-192.