NEWS RECOMMENDING BASED ON TEXT SIMILARITY AND USER BEHAVIOUR

Dušan Zeleník, Mária Bieliková
Institute of Informatics and Software Engineering, Faculty of Informatics and Information Technologies, Slovak University of Technology, Ilkovičova 3, 842 16 Bratislava, Slovakia
zelenik@fiit.stuba.sk, bielik@fiit.stuba.sk

Keywords: recommendation, personalization, behaviour, monitoring, similarity, news web portal, news, readers.

Abstract: In this paper we describe a method for recommending news on a news portal based on our novel representation by a similarity tree. Our method for recommending articles is based on their content. The recommendation employs a hierarchical incremental clustering which is used to discover additional information for effective recommending. The important and novel part of our method is an approach to discovering the interests of individual readers using tree structure created according to similarity of articles. We concentrate on enabling the recommendations in any time, i.e. we discover user's interests real-time. Our method discovers specific interests of the reader using information gained from monitoring his activities in the news portal. We describe the mechanisms for recommending up-to-date and relevant articles. It is based on known solutions, but incorporates unique representation of user interests by binary tree. Moreover, our aim was to provide recommendations in real-time. Recommendations are thus generated depending on the actual reader's interest. We also present an evaluation of recommendations in the experiment where we use accounts of real readers and their history of reading.

1 INTRODUCTION

Making personalized recommendations is nowadays becoming increasingly popular topic. There are several reasons for this. The main reason is the size of information space containing sources to be recommended and the inability of a human to browse this space in full in order to find relevant information. Our main concern is to facilitate the exploring activity of the user through the information space by recommendations with stress on changes of the user interests in time. Recommendations are also a perfect tool for marketers. Targeted advertising is linked to the analysis and consumer needs. Thus, especially in e-shops and web business, recommendation of goods commonly takes place.

Area of news is a typical example of a comprehensive information space. Online newspapers aim at keeping their readers interested. They dedicate effort to search for improvements, in particular, to bring comfort. Amounts of articles which are added daily are thus processed to be recommended to users according to their needs.

In the field of news, we should consider time sensitivity of articles. We can expect that our content changes dynamically and our users change their interests in time too. Recommending articles is then time sensitive and should be done real time to preserve relevancy of the news.

In this paper we describe our proposal of content-based recommending. Recommendations are made based primarily on a history of the reading. For recommendation decision we use the individual user activity (recent articles read) to predict the content he is interested in. We choose content-based recommender due to the increasing opportunities in the processing of content. Besides, regarding news, content is definitely important and valuable.

Our method for news recommending uses incremental hierarchical clustering. Clustering is carried out using textual similarities and is adapted to allow rapid up-to-date and personalized recommendations.

As a part of the sme.fiit project (Barla et al., 2010), which aims at news recommendation in largest electronic Slovak newspaper (www.sme.sk) we present in this paper a recommender called TRECOM. TRECOM is based on monitored user behaviour and processed news articles represented effectively considering the news similarity. Data used for the recommendation are taken from the web site SME.sk and contain news texts and the user history (logs of news reading). We employ a reader's activities history from the news portal without any feedback, but the intention is similar to the one presented in the related work (Carvalho et al., 2005) where web logs and user history is used.

2 RELATED WORK

Generating personalized recommendations is related to the observation of a single user behaviour followed by items suggesting using either content-based or collaborative filtering methods (Su and Khoshgoftaar, 2009). In advance there are also hybrid techniques, which could be used to recommend items (Burke, 2002). These methods often combine both principles to avoid negative aspects in both types.

Collaborative filtering methods for news recommending are based on presumption that the majority of similar users have found something interesting for the rest (Suchal and Návrat, 2010). Actually, this is more about predicting the behaviour than about discovering the interest or needs. This approach has one advantage in comparison to content-based approach. We are able to surprise the user and keep the relevancy of the article in the same time (Ge and Delgado-battenfeld, 2010).

Mooney (Money and Roy, 2000) proposed a method for book recommendation where each text is processed and represented using text categorization. They claim that content-based recommenders are best at recommending unpopular items when there is not sufficient information about users, but content information is easy to obtain. In our case, we have news relatively easy to process.

There are more options how to calculate similarity for texts. As it was mentioned, in related work (Tintarev and Masthoff, 2006) even simple solutions like Bag Of Words are accurate enough for news recommending, considering the fact that it has low complexity. Complexity is important if we want to provide real-time recommendations.

There is always a problem with unknown or new users (Adomavicius and Tuzhilin, 2005). These users have not been monitored to enable any recommender to estimate their interests or needs. Typical solution is to recommend random, the newest or the most popular items.

Serious problem is also overspecialization which happens especially in content-based recommenders. This could be solved by randomly generated recommendations and omitting the items which are very similar to those which the reader already saw like in DailyLearner (Billsus and Pazzani, 2000).

There are also other aspects of user state which should be considered when we want to recommend news. As it was mentioned in paper on news recommending (Jancsary et al., 2010) there are context-sensitive features. To involve these aspects we need to find them in real-time. Respectively, we have to affect the recommendations in real-time.

3 NEWS PORTAL

We recommend articles which are at the web-based news portal. We have to face time sensitivity, variety and amounts of articles and readers. Our users are readers of this news called SME. There are

- around 350 thousands of visits every day;
- authors add around 250 new articles every day in 430 different categories (combination of category and section);
- average user reads 2 articles per a day and spends almost 17 minutes at this site a day.

Articles comprise information about the time of publication, author, section, category and more. Time of publication is important attribute. It defines time sensitivity for this domain. Old articles lose importance over time, despite of the relevance of their content for specific user. Generally we have to find personalized and the most recent articles.

For recommending recent articles in dynamic environment of news we should use a representation of articles which allows incremental adding articles. Retrieving articles and searching user interests has to be based on algorithms with low complexity to be able recommend in real time with preserving recency of the news. Besides mentioned time sensitivity we should reflect changing user interests.

Readers of the news do not want to be overwhelmed by the same information. There is a need to vary articles which are recommended. The reader gradually uses the recommendations, so it is appropriate to vary these recommendations over time. Not because of the news recency only, but also because of the changing or deepening user interest. Constrained list of recommended items should cover majority of momentary user's interests.

4 METHOD FOR NEWS RECOMMENDING

The first phase is to discover the interests of individual users. This is done by monitoring the activity of each reader. Articles that readers display are located in a hierarchical structure we designed. This structure keeps relations between similar articles. We discover user interests using the records of user activity and the hierarchy. We describe the way how to locate articles that are appropriate to the reader. Another task is to compile list of articles. The number of recommendation should be constrained to a limited number. Therefore, we need to find the equilibrium between recency and relevancy to maximize precision.

4.1 Discovering Interests

A prerequisite for our method is that each individual has some interests. This can be easily verified using a history of readings of particular readers. We can follow the interest in certain categories or sections of news portal. Figure 1 presents the selected reader and records of his activities during the period of 15 days. We can see that the interest of certain categories prevails over the others.

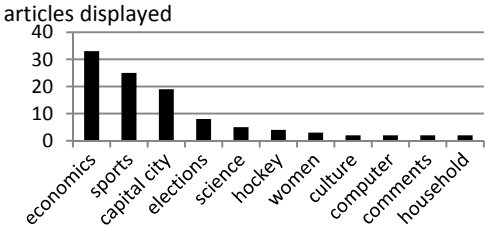

Figure 1: Top (of 40) categories displayed by the reader.

Similarly, there are identifiable fields of interests for each reader. It makes sense to explore more interests based on the calculated similarity between articles. We substituted this metadata (categories) made by editors by the hierarchy of similarity relations which provides its own metadata.

We use a hierarchy of relations, which is incrementally built, similarly to the hierarchy presented by Sahoo (Sahoo et al., 2005). In our hierarchy, we can rely on the repository which contains current articles and also assume that they are properly organized. Set of words extracted from articles and normalized are used as features to compute similarities among articles. Each node in the tree is labelled by a set of features. Edges in the tree represent the hierarchy which keeps similar articles nearby. We designed our representation as a hierarchy where

- real articles are placed at the lowest level of the tree (leaf nodes),
- features are spread to the meta level of the structure,
- similarity is kept in the hierarchy.

There are several options how to calculate similarity based on the content itself. There are also sophisticated methods, which are able to determine semantic similarity (Gabrilovich and Markovitch, 2007). However, simple text similarity is often used in news recommending and it gives good results (Kroha and Baeza-Yates, 2005). We use Jaccard's similarity to calculate articles similarity.

Figure 2 shows a way of discovering the reader's interests. We use the tree structure created using the similarity of articles and records of user activity. We have a hierarchy of nodes which effectively represent similarity of real articles even without actual calculation between particular pairs. Thick edges are paths from the displayed article to the root of the tree. Nodes where are thick edges merged are fields of interests.

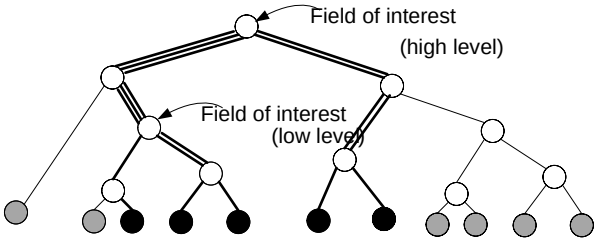

Figure 2: Discovering interests. Black nodes represent already displayed articles.

In this manner, we discover interests for each user using his history of reading. One user interest is one node in the tree which is used to define the set of articles belonging to this interest (articles in the subtree). Since we use a tree structure we work with hierarchy of these interests.

4.2 Retrieving Suggestions

When considering reader's fields of interests we can compare fineness of the interests depending on the depth where the node of the tree is located. Fields of interest that are closer to the leaves of the tree are more focused on a particular topic (e.g. articles about hockey). Fields of interests that are closer to the root are dedicated to more general topics (e.g. articles about the sport). This structure has some useful properties.

Recommending specific articles using our proposed hierarchical structure is a matter of selecting articles from more interests. We find the relevant interest for a particular reader. The relevance of the interest is calculated as the ratio of articles displayed from this interest and all articles belonging to the interest.

Thus, we are able to sort interests and prepare for the selection of appropriate articles. There are obviously plenty of appropriate articles, since rapid growth of the news dataset (250 new articles per day). Figure 3 presents the selection of interesting articles for a specific reader.

Highlighted articles are those which the user has read. These articles are used to determine the relevance of his interest. Other articles are potentially interesting for the reader. We selected the articles, which are the subject of further recommendations.

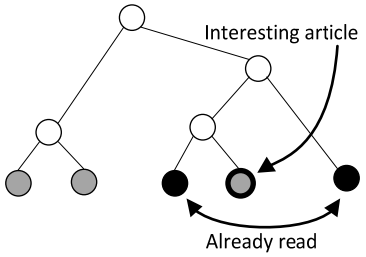

Figure 3: Selecting interesting articles.

Because of the need to avoid overspecialization, we penalize very similar interests. Otherwise, recommended articles would have been closely similar to those already displayed by the reader.

4.3 Compiling Recommendations

The reader has sometimes problem also with long list of recommendations (Bollen et al., 2010). Therefore, we choose articles that cover all relevant interests of the reader but are from distinct fields.

We also integrate the time as an attribute in the compilation of the list of recommendations. Time is an important attribute, which could indicate whether the interest that we discovered is outdated or not. We introduce the additional information that is maintained in a hierarchical structure. We can find the latest article which was added for each branch of the tree. Time attribute is spread as maximum of two sub-branches. This way we can efficiently identify the most recent article and the time when it was published. Interest is then as relevant as the last added article. We gain the possibility to combine time relevancy and the content relevancy. To create a list of recommended articles we considered both attributes. The method is described in the following steps and Figure 4.

1. Selection of articles displayed by a reader
2. Discovery of areas of interests in the tree
3. Selection of unread articles for each interest
4. Sorting articles by time in particular interest
5. Creation of a matrix containing interest
6. Linking the columns of the matrix into a list as illustrated in Figure 4.

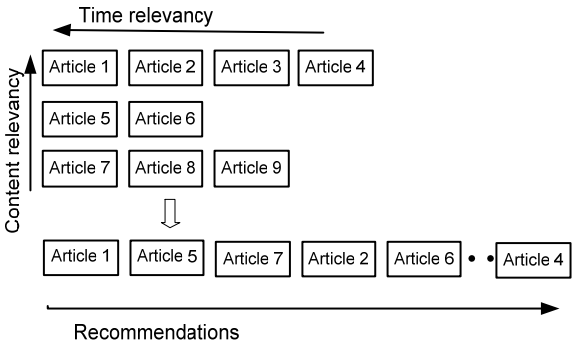

Figure 4: Compiling the mix of recommendations. Articles in rows belong to the same field of interest. The most relevant interest is at the top. The most recent articles for each interest are on the left side. Articles 1-9 are formed into list of recommendations by columns.

The list covers interests of the user. Our method is designed to recommend 10 items for one request to avoid choice overload (Bollen et al., 2010). Articles are not only from one theme but cover more topics. Articles are even up-to-date.

We are able to create the list in every moment and in real-time. This is mainly because of the hierarchy which we use to represent relations between articles. Operations are fast enough to generate the matrix of recommendations. Actually, we change the whole matrix only in cases where the user initiates a new session, or exhausts the list.

5 EVALUATION

We performed several experiments conducted in the real environment of the news portal. This brings real data as articles, readers and their activity (thousands of articles and readers).

One way of evaluation, we made, is the real usage of the method and a user feedback. We selected few readers and formed the controlled group. This group had to evaluate recommended articles and become familiar with our recommender.

Users have had a chance to use our recommender through the browser plug-in. This plug-in works as an extension to the news portal. We enriched the news portal with a list of recommendations. We also added a simple voting control to each article only for the evaluation purposes.

The experiment was conducted with 10 people who rated 88 recommended articles during ordinary reading. Readers evaluated recommended articles using binary values (appropriate, inappropriate). Reader had a list which was changed every hour. It was not obligate to evaluate every article from the list. We took 88 rated articles and 62 articles were positively evaluated. Our accuracy with this controlled group was 70%. It means that 70% of recommended items covered interests of readers.

Our second experiment was based on synthetic tests. In this case we simulated the feedback received from readers on the basis of their actual behaviour in past. Since we are talking about simulating the evaluation, we use many more readers than in previous experiment. The entire test was executed with a set of 1, 000 active readers and their reading records (5 days, 20 articles per day for average reader).

We divided the records from complete history of readers into two smaller intervals. The first interval is denoted as training interval and second as a test interval. The first interval is used to generate recommendations using our method. It is the same list of recommendations which would appear when using the website at the end of this period. Real history in test interval is then compared with recommendations generated using training interval.

However, the reader displays thematically similar articles to the recommended articles. To examine whether recommendations cover reader's interests or not, we did not compared exact articles. We compared articles using the similarity. We did not use our relations to compare articles. We used pair of section and category provided by the news website to be objective.

We used 1,000 active readers and their history. To be accurate, there are around 430 valid combinations of sections and categories. It means there are around 430 options to pick correct combination.. We evaluated if the recommendation is the same combination of section and category as the article in testing period.

Figure 5 indicates the precision and recall for more testing intervals. We can compare the length of the intervals used to calculate the recommendations. We see that the precision is growing up to 60%. The recommendation is correct in 60% of the cases.

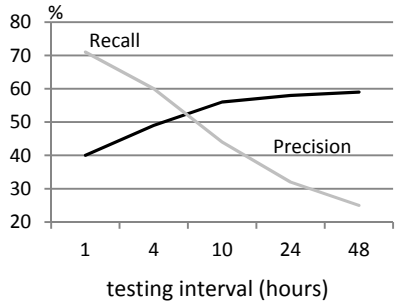

Figure 5: Precision and recall plotted in the chart.

To make a better picture we have compared our results with results of the other content-based recommender which uses the same dataset and evaluation method (Kompan et al., 2010). We have observed that our method has significantly higher recall for shorter testing intervals. Our recommender was able to cover user's interests also in 1 hour interval. This happens because our recommender uses composition to cover as much user's interests as it is possible.

From a user's perspective it is often important to know the method for recommending and how these recommendations are calculated (Ahn et al., 2010). Our user is willing to accept advice if he knows how the machine discovered this advice. We discovered that sometimes just an outline of the solution helps.

We found out that we are logging also articles already recommended. These records are used again for recommending. Discovering of interests is inappropriate when the calculation works with these articles. In fact, recommenders should not replace the standard navigation, but they should satisfy the user with an additional functionality. Otherwise, the information space may be undesirable narrowed by these recommendations. One solution could be the addition of random articles, which would allow the user to navigate into these hidden areas.

We also observed the problem with non-active readers. This means that the interval used for recommending was not sufficient to discover interests. We used only 5 days of the history of reading. The average user reads only two articles per a day. Shorter intervals are not sufficient because our method is not able to discover enough interests.

6 CONCLUSIONS

In this paper we described a method for personalized news recommending. We focused on the content of articles and the user's interests. We used an effective

representation of the similarity relations between the articles. Advantages of this hierarchy include logarithmical complexity, metadata which are generated using content of the articles and incremental approach. This is useful if we need real-time calculation and the metadata provided by the authors of the news are not sufficient. On the other hand, a disadvantage is that the tree structure could not provide relations which are not transitive (i.e. text similarity of news).

We use properties of the hierarchical representation in our method. The results thus meet the requirements of the recommender system. Hierarchical clustering has low, logarithmical complexity of storing and retrieving articles. The hierarchy enables us to discover interests for every moment using the history of reading.

Our main contribution is utilization of hierarchical structure, which incrementally generates metadata about articles. Meta-documents which are created this way have inheritance relations. These relations represent similarity between real articles. The advantages of our recommender systems are linked to this representation. We are able to discover user's interests in real-time, even if we use vast information space to recommend news.

We focused in our work on real-time content-based recommending. Our future work includes considering the context of the user's interests. We plan to improve our recommender to consider the actual interests of a user. We have a presumption that interests change in time, with location, mood or emotions. Since we are able to recommend news in real-time, this is mainly a matter of recognizing the behavioural patterns and contexts.

ACKNOWLEDGEMENTS

This work was supported by the Scientific Grant Agency of SR, grants No. VG1/0508/09 and VG1/0675/11, and it is a partial result of the Research & Development Operational Program for the project Support of Center of Excellence for Smart Technologies, Systems and Services II, ITMS 25240120029, co-funded by ERDF.

REFERENCES

Ahn, J., Brusilovsky, P., Grady, J., He, D., and Syn, S. Y. 2007. Open user profiles for adaptive news systems: help or harm*?*. In *Proc. of the 16th int. Conf.on World Wide Web. WWW '07*. ACM, New York, NY, 11-20.

Adomavicius, G. and Tuzhilin, A. 2005. Toward the Next Generation of Recommender Systems. IEEE *Trans. on Knowl. and Data Eng*. 17, 6, 734-749.

Barla, M. et al., 2010. News recommendation. In *Proc. of the 9th Znalosti*, Jindrichuv Hradec., 171-174.

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

Bollen, D., Knijnenburg, B. P., & Graus, M. 2010. Understanding Choice Overload in Recommender Systems Categories and Subject Descriptors. In *Proc. of 4th ACM Conf. on Recommender Systems*. Barcelona, Spain, 63-70.

Burke, R. 2002. Hybrid Recommender Systems: Survey and Experiments*. User Modeling and User-Adapted Interaction* 12, 4 (Nov. 2002), 331-370.

Carvalho, C., Jorge, A. M., and Soares, C. 2006. Personalization of E-newsletters Based on Web Log Analysis and Clustering. In *Proc. of the IEEE/WIC/ACM Int. Conf. on Web intelligence. IEEE Computer Society*, WDC, 724-727.

Gabrilovich, E., Markovitch, S. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proc. of the 20th int. Joint Conf. on Artificial Intelligence*, Hyderabad., India, 1606-1611.

Ge, M., Delgado-battenfeld, C. 2010. Beyond Accuracy: Evaluating Recommender Systems by Coverage and Serendipity. In *Proc. of 4th ACM Conf. on Recommender. Systems*. Barcelona, Spain, 257-260.

Jancsary, J., Neubarth, F., Trost, H. 2010. Towards Context-Aware Personalization and a Broad Perspective on the Semantics of News Articles. In *Proc. of 4th ACM Conf. on Recommender Systems, Barcelona,* Spain, 289-292.

Kroha, P.,Baeza-Yates, R., 2005. News classification based on term frequency. In *Proc. of the 16th Conf. on Database and Expert Sys. Apps*, 428–432.

Kompan, M., Bieliková, M., 2010. Content-Based News Recommendation. In *Proc. of the 11th Conf. EC-WEB. Springer-Verlag, Bilbao, Spain,* 61-72.

Mooney, R. J. and Roy, L. 2000. Content-based book recommending using learning for text categorization. In *Proc. of the 5th Conf. on Digital Libraries*, TX, USA, 195-204.

Sahoo, N., Callan, J., Krishnan, R., Duncan, G., Padman, R. 2006. Incremental hierarchical clustering of text documents. In *Proc. of the 15th ACM int. Conf. on Information and knowledge management*, NY, USA, 357-366.

Su, X. and Khoshgoftaar, T. M. 2009. A survey of collaborative filtering techniques*. Adv. in AI, 36-55.*

Suchal, J., Navrat, P. 2010. Full text search engine as scalable k-nearest neighbor recommendation system. In *Proc. of the AI in Theory and Practice 2010*. WCC. IFIP AICT 331, Springer, Boston, 165-173.

Tintarev, N., Masthoff, J. 2006. Similarity for news recommender systems. In *Proc. of the AH'06 Workshop on Recommender Systems and Intelligent User Interfaces, 1-8.*