Virtual Community Detection in Vast Information Spaces

Marián HÖNSCH*

Slovak University of Technology
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
xhonsch@stuba.sk

Collaborative filtering is present within current web-based systems in many forms. At the beginning they were mostly either item based or user based, but as the time passed, many hybrid approaches combining several techniques from multiple disciplines emerged. However, the basic idea remained always the same: use past experiences of users to get benefits for an individual.

We analyzed research works related to collaborative filtering in domains of research papers recommenders, personalized learning systems, news recommenders etc.. Used forms of collaborative filtering differ in ways how they model the user, gather users opinions, compute similarity between users or items, define groups of users, present recommendations and many others. We will consider these observations and also other features in our work. As collaborative filtering is always tailored to the specific domain, we chose to focus on recommending news articles. Our main aim is to develop a method for detecting virtual communities among users in order to improve recommendations on a news web portal.

Virtual community detection is an advanced feature of recommender systems, built on the top of basic recommenders. We need to ensure at least a stable method of user modelling and similarity counting. In our case, user model is based on keywords, as we assume that the interests of a user can be projected into keywords extracted from content he reads [2]. To handle synonyms and ambiguous words present in a user model we need to employ additional knowledge about keywords. One approach is to build our own semantic model by processing all the available articles and tracking what words do appear together, another one is to use a semantic service such as WordNet. To handle ambiguous words we use other words from the same article, which help us to reveal the right sense. A much simpler alternative to semantically-based approaches is to do a word count.

Similarly to the representation of user interests, an article is also represented by keywords extracted from it (keyword-based domain model). When a user reads an

* Supervisor: Michal Barla, Institute of Informatics and Software Engineering

Spring 2010 PeWe Workshop, April 18, 2010, pp. 45–46.

article the domain model (keywords of the article) is interlinked with the user profile. Keyword model of an article is constructed mainly from headlines and words from the first few sentences. We assume that these parts contain most relevant content.

In a user profile we gather all information that the system knows about the user. There can be many sorts of attributes and inputs. We model these attributes in layers (i.e., synonyms, accepted recommendations, negative feedback, long term information). A crucial task of every recommender system is to gather user opinions. In our system we want to consider feedback in a form that user has read an article, did or did not accept recommendation. To handle a "cold start" problem and the overload of a user profile we fade and forget items. Then two users would have profiles of comparable size, independently of how long they have been actually using the system.

A community is formed by people with similar user models. We cluster user profile so that every cluster represents an interest [3]. This approach is based on the assumption that different categories of articles are represented by different clusters of keywords. The categories correspond to the users interests. Then we group users based on these partitions. Community is represented by an aggregated keyword model extracted from user models of its members. Communities tend to change radically over time and appropriate reaction of the system to these changes is an actual research field. In [1] they propose methods how to count virtual communities with respect to changes over a short time. A newspaper domain is a good example. Only rarely people read the newspaper from yesterday. People also tend to read different categories of articles depending on the time of the day or day of the week (i.e., on weekends we read the Sunday part). Detected communities are volatile so we compute new compound of communities on a daily basis.

To test our approach we recommend articles. We compare the quality of our recommendations against existing collaborative filtering approaches. There are two types of recommendations, one is *novel items recommendation* (i.e., what others read before) and *suggestions to the article the user is actually reading* (i.e., other that read this have also read this). Our main contribution is detecting virtual communities based on user profile clustering, layered user model and community graphs. We deal with fluctuating and time-dependent changes in community detection.

Acknowledgement. This work was partially supported by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. KEGA 028-025STU-4/2010.

References

[1] Falkowski, T. and Spiliopoulou, M. 2007 Users in Volatile Communities: Studying Active Participation and Community Evolution. User Modeling 2007, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 47-56.

[2] Lops, O., Degemmis, M. and Semeraro, G. 2007 Improving Social Filtering Techniques Through WordNet-Based User Profiles. User Modeling 2007, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 268-277.

[3] Zhang, M. and Hurley, N. 2009. Novel Item Recommendation by User Profile Partitioning. In Proc. of the 2009 IEEE/WIC/ACM international Joint Conference on Web intelligence and intelligent Agent Technology - Volume 01