Detecting User Communities Based on Latent and Dynamic Interest on a News Portal

Marián Hönsch*
Institute of Informatics and Software Engineering
Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava
Ilkovičova 3, 842 16 Bratislava, Slovakia
marian.honsch@gmail.com

Abstract

This paper describes our work on identifying communities of individuals based on their interests while browsing the web. A user can belong to several communities at a time, where each community represents parts of his interests. We assume that recommendations coming from such communities are more accurate than from communities based on a whole user profile. We describe how to record and identify particular interests for each user. Interests evolve from analysis of the resources that the user has viewed in the past and are defined as cluster of keywords. Based on the time period we model short term and long term interests. The novel approach is to create virtual communities based on these interests, both short and long. To evaluate our approach we built articles recommender for a news portal. As recommender systems are tailored to the specific domain, we also adapted our approach slightly to better fit the news portal domain, which is highly dynamic and with frequent changes. We consider these time-dependent changes by weighting the influence of volatile communities on recommendations.

Categories and Subject Descriptors

I.7.5 [**Document and Text Processing**]: Document Capture—*Document analysis*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*User profiles and alert services*; H.3.5 [**Information Systems**]: Information Storage and Retrieval—*Online Information Services*

*Master degree study programme in field Software Engineering. Supervisor: Doctor Michal Barla, Institute of Informatics and Software Engineering, Faculty of Informatics and Information Technologies, STU in Bratislava. Work described in this paper was presented at the 7th Student Research Conference in Informatics and Information Technologies IIT.SRC 2011.

Hönsch, M. Detecting User Communities Based on Latent and Dynamic Interest on a News Portal. Information Sciences and Technologies Bulletin of the ACM Slovakia, Special Section on Student Research in Informatics and Information Technologies, Vol. 3, No. 2 (2011) 47-50

Keywords

virtual communities, user interests, recommendations, latent analysis, semantic keyword relatedness

1. Introduction

One of the drawbacks of today's community-based collaborative recommender systems is that they group users based only on their aggregated similarity. Only those users are assigned to a community, whose profiles match completely the profile of the community. This prevents them from using the wisdom of the crowd coming from users that match only parts of their interests. This negative impact was also described in former research papers [4, 9, 3, 1]. We assume that recommendations coming from communities that address this issue can significantly improve the quality of recommendations. The demand for accurate recommender systems is very actual. Not only that users prefer personalized portals over non personalized ones, but internet companies can also raise their profits by having users spend more time on their web page thanks to the recommender and personalisation features.

In our work we address the negative impact of aggregated similarity on user grouping by proposing a method consisting of two main steps:

1. Collect and detect users interests
2. Create relatedness graph between interests and detect virtual communities

Our target is to find communities that are defined by one particular interest. This community should include all users that share this interest. Depending on how many interests we have discovered in his profile, he can belong to several communities. Interests are derived from the analysis of the domains corpus and are expressed as sets of words that have dense relatedness interconnection between each other. Next we cluster the interest based on our similarity metric and so detect communities.

We evaluated the whole approach on a news recommender system by recommending articles based on communities detected with our approach. We also performed an experiment to confirm our interest comparing strategy and compared it to the approach described in [7].

The paper is structured as follows in second chapter we provide an overview of existing article community-based

recommenders and we mention algorithms to detect virtual communities. We take a close look on how to detect and maintain user interest based on his actions in the third chapter. In fourth chapter we describe how to form communities based on detected interests and in the fifth chapter we describe the interest comparing strategy. In the last chapter we present our actual results.

2. Related work

The area of recommending to communities is widely covered in research papers and there are also many portals that adopt the gained knowledge. We focused on existing solutions of recommender systems for communities on news portals. One of the most common is Google News[1], which groups users according to their geographical position, the context of their visit [4]. They include several news portals but do not analyse the content of the news articles.

Grouping of strongly related articles based on content or user behaviour analysis can serve for storytelling and breaking news detection, which can also lead to improvements of recommendation quality and better prediction of user behaviour [10].

Mobile devices seem to be a suitable platform for recommendations, as they deal well with personification of user accesses and context gathering problems. Authors in [12] recommend relevant articles by means of a personalized news selection on a mobile device. Their approach is based on extracting different topics in form of keyword sets from an article.

In the work described in [6] the authors model a document as a graph of semantic relationships between terms of that document. On that graph they applied the algorithms to detect virtual groups. A detected group is a specific group of words considered to characterize a particular topic from the document. Their semantic relatedness measure between two words is derived from their co-occurrence in articles of Wikipedia, also referred as the Wikipedia-based semantic relatedness [6].

There are lots of described and tested algorithms in the field of virtual communities detection, each targeting a certain domain or graph structure. See [5] for a good overview over state-of-art algorithms. In our work we cannot strictly decide which algorithm to use, because this domain is not fully researched from this perspective. That is why we decided to apply several of them and join their results using appropriate weighting mechanism.

3. Identifying user interests

An interest of a user is deducted from the articles he has accessed before. Based on the time interval and frequency of accesses we can deduce long term interests from short term interests. Long term interests are such short term interests which are repeatedly detected over time. We assume that keywords extracted from accessed articles determine users interests. Similar assumption can be found in [12] and [6]. Our interest extraction process can be summarized in these steps:

1. Capture all articles accessed by a user.
2. Extract keywords from these articles.
3. Create keyword relatedness graph by connecting the keywords based on their relatedness (i.e. semantic relatedness)
4. Find virtual communities based on the keyword relatedness graph.

Detected keyword groups are the base of user interests. With a similar approach on all articles we can detect interests in the whole corpus. We refer to users interests as *local* interests and to interests detected in corpus as *global* interests. Local interests have lesser cardinality than global interests and they are subsets of global interests. As a keyword relatedness, to which we refer in step three, we can use the word relatedness from dictionaries or thesauri such as WordNet[2], which was successfully applied in [9], or the relatedness gained trough latent analysis, which was successfully applied in [6] and [8]. In our experiment we propose to use the Latent Dirichlet Allocation - LDA on the domain content and derive the semantic relatedness from its results. This makes the detected groups *latent* as the relatedness is depended on the LDA analysis and is influenced by the analysed content.

The detection of interests is repeated in subsequent time intervals in order to keep pace with the dynamic of a news portal. The results obtained from previous iterations of the detection are considered in the actual interval with a certain factor. Interests that appear for one user in many recent intervals are regarded as *long term* interests and only recently discovered interests are considered as *short term* interests. To prevent overfitting and overspecialisation [1] we reduce the influence of past accessed articles. The older results gain less weight, which means that if an interest is not periodically strengthened it will fade out. On the other hand, by considering older results we reduce the influence of new and strong interests that can be, however, only temporal.

To capture user activities we use the overlay user model, which was proved to be suitable for user modeling in an open corpus domain [2]. Our corpus consists of keywords, which were extracted from all articles. This corpus is constantly extended as new articles are added. The user model is keyword-based, it lays over the keyword corpus. A user model contains words, which were extracted from articles he has accessed before.

4. Detecting communities

A community includes users who have one particular interest in common. To find neighbours for one of users interests we compare it to all other identified interests. An interest is a set of words with dense semantic relatedness connections. To compare it with other interests we compare all these sets. The neighbours are then all interests that match to a certain threshold. We define two strategies for interest comparison:

- local interest to local interest
- local interest to global interest

[1] Google news: http://news.google.com/

[2] http://wordnet.princeton.edu/

The first approach detects all users that have a similar interest to the selected user interest and therefore it is more specific for one user, compared to the second approach which groups interests relevant to a specific global interest. The community size is cut-off by a threshold or top N rule. First method is similar to finding communities with activation spread [5]. Second approach groups users according to global interests. A global interest is static and all users need to have a certain matching threshold to be considered as members of a community determined by this interest. Such communities would not vary based on one user. We can also refer to it as clustering based on common properties.

A certain boost to both approaches is to consider the article category specified by the news editor (i.e. Sport, Politics) from which the keyword was extracted.

To ensure the uniform cardinalities of discovered communities we split big communities to smaller ones and merge smaller communities together. This can be achieved by changing the match threshold or by merging and splitting global interests. We either split the interest by repeated searching for virtual communities within it, as described in step four in chapter 3, or connect it with the other and most similar interest.

Communities are detected continuously, each time when computed interests significantly change. This can be a period between few hours to several days. It also depends on the users activity. Here we apply the same technique of long term and short term communities as described on long term and short term interests in chapter 3. We consider the past results in the actual time interval, but the influence will fade as the time passes. This also reduces the impact of strong short term interests like actual social events or happenings.

5. Comparing interests

By comparing two interests we aim to determine the relatedness between them. The interests are represented as a set of words. Therefore to compare them we must perform a comparison of two sets of words. As basis we have chosen to use two metrics:

- exact match
- relatedness match

Exact match is a comparison of words for exact match between them. This method resembles to "bag of words" text operations. This method was successfully applied on news articles in [7].

Relatedness match considers weighted connection we have introduced by creating the keywords relatedness graph. We search for the shortest path between two words in the keyword graph. The result value is the average of sums of the weights from the shortest path for each pair of words from both sets. Below is the formula for similarity of two sets.

$$\mathrm{sim}\,(set_1 \mid set_2) = \frac{\sum_{i=0}^{size(set_1)} \left(\frac{\sum_{j=0}^{size(set_2)} weight_{w_i,w_j}}{size(set_2)} \right)}{(\log_2 size\,(set_1) + 1) * 3} + \frac{\sum_{j=0}^{size(set_2)} \left(\frac{\sum_{i=0}^{size(set_1)} weight_{w_j,w_i}}{size(set_1)} \right)}{(\log_2 size\,(set_2) + 1) * 3}$$

Where $w$eight x, y means the semantics relatedness between words x and y. We use the logarithm with base 2 to normalize the size of sets. This process is a generic comparator of two sets of words (i.e. we also compare articles with it).

6. Evaluation

We conducted our experiment on data acquired from a Slovak news portal sme.sk[3]. In order to achieve more accurate results we employed natural language processing tools (filtering, lemmatising) and keyword extraction[4]. For the LDA we used open source libraries provided on Google Code[5].

As input we used the log of accesses of users to articles capturing a longer period of time and the actual content of these articles. We performed keyword extraction, created the keyword relatedness graph based on the LDA result and identified local and global interests. During this process we processed 60 650 articles and extracted 180 588 unique words.

We present results of articles comparison, which uses the same core concept as our interest comparison method. We employed the same dataset and evaluation metric as in work [7]. Their method is similar to the exact match metric of our solution. The manually annotated similarities of article pairs were defined as related and nonrelated. We used our approach to determine similarity for every pair of articles. We evaluated the success rate of correctly classified articles pairs as related and nonrelated. The results are listed in Table 1. We slightly outperformed the method in [7]. This is probably based on our added relatedness measure.

Table 1: Experiment results of comparison.

	Recall	**Precision**	**F meas.**
Selecting all related pairs	*0,780*	*0,951*	*0,856*
Selecting all unrelated pairs	*0,999*	*0,993*	*0,995*
Results presented in [7] selecting all related pairs	*0,700*	*0,816*	*0.753*

We performed recommendation evaluation as a predictor. Recommendations come from other community members

[3]SME: http://www.sme.sk/
[4]METALL: http://peweproxy.fiit.stuba.sk/metall/
[5]PLDA: http://code.google.com/p/plda/

Table 2: Experiment results of recommendation.

Removed top N articles	Our Method		Random Groups		Results from [11]	
	Precision	Used subset	Precision	Used subset	Precision	Used subset
0	31,2%	19,6%	27,7%	10%	25%	-
10	27,6%	35%	25%	12,4%	-	-
20	19,9%	40%	15,9%	16,5%	-	-
30	17,8%	42%	12%	20,7%	-	-
40	15 %	45%	9,5%	21%	-	-

using collaborative filtering to find the most interesting unseen content for the particular user. We generated recommendations based on short term interests. Our training phase was a one hour period and we tested it on the following hour. During the experiment we observed, that the groups have strong tendencies to recommend articles listed in the top N list of most accessed articles during last hour. We generated mostly same recommendations for all users and recommended only a $\sim 20\%$ subset of all articles. We executed also the experiment with randomly generated groups. The random groups led to recommendations, which were in 85% of cases identical to the top 20 list. Our approach delivered nearly same results as random groups. Therefore we repeated the experiment and excluded the articles listed in the top N list from final recommendation. After this modification our approach used a $\sim 40\%$ subset of all articles and outperformed results obtained with randomly generated groups.

We executed our experiment on the same sme.sk dataset as described in [11]. We have chosen this way of evaluation to be able to compare with their results. In the Table 2 we list comparison of achieved precisions for each approach (our, random one and theirs). We also show the percentage of used subset of articles. The rows present amount of excluded articles from recommendations based on the top N list.

7. Conclusions

In this paper, we presented a new method of grouping users based on their individual interests and its basic evaluation on data from a news portal. Our main contribution is the novel approach to user interest deduction and the detection of virtual communities based on these interests. In our work we introduced latent analysis of content the user has accessed to reveal relationships between words, which form the base for further processing of interests. The content analysis by LDA requires a larger amount of input data to generate a sufficient domain model, but we see a substantial advantage in the fact that the deducted relatedness is based on domain data and is therefore more suitable.

We also handle the dynamics of a news portal environment by repeating the computation of interests over time and by pro-actively maintaining the gathered user profiles. This allows us to find long term and short term communities, which can lead to better understanding of user behaviour and community dynamics. By considering previous results in current iteration we introduce a model to lower the influence of strong interests that can be temporal and strengthen long term interests.

Acknowledgements. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0675/11.

References

[1] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005.

[2] M. Barla. Towards Social-based User Modeling and Personalization. *Information Sciences and Technologies Bulletin of the ACM Slovakia*, 3(1):52–60, 2011.

[3] P. T. Braak, N. Abdullah, and Y. Xu. Improving the Performance of Collaborative Filtering Recommender Systems through User Profile Clustering. In *Web Intelligence & Intelligent Agent*, pages 147–150, Washington, DC, USA, 2009. IEEE Computer Society.

[4] A. S. Das, M. Datar, A. Garg, and S. Rajaram. Google news personalization. In *Proceedings of the 16th international conference on World Wide Web - WWW '07*, pages 271–280, New York, New York, USA, 2007. ACM Press.

[5] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, Feb. 2010.

[6] M. Grineva, M. Grinev, and D. Lizorkin. Extracting key terms from noisy and multitheme documents. In *Proceedings of the 18th international conference on World wide web - WWW '09*, pages 661–670. ACM Press, 2009.

[7] M. Kompan and M. Bieliková. Content-Based News Recommendation. In *E-Commerce and Web Technologies: 11th Int. Conf., EC-Web 2010, Proc.*, volume 61 of *Lecture Notes in Business Information Processing*, pages 61–72. Springer Berlin Heidelberg, 2010.

[8] R. Krestel and B. Mehta. Learning the Importance of Latent Topics to Discover Highly Influential News Items. In *KI 2010: Advances in Artificial Intelligence*, volume 6359 of *Lecture Notes in Computer Science*, pages 211–218. Springer Berlin Heidelberg, 2010.

[9] P. Lops, M. Degemmis, and G. Semeraro. Improving Social Filtering Techniques Through WordNet-Based User Profiles. In *User Modeling 2007*, volume 4511 of *Lecture Notes in Computer Science*, pages 268–277. Springer Berlin Heidelberg, 2007.

[10] M. Mori, T. Miura, and I. Shioya. Topic Detection and Tracking for News Web Pages. In *2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 338–342. IEEE, Dec. 2006.

[11] J. Suchal and P. Návrat. Full text search engine as scalable k-nearest neighbor recommendation system. In M. Bramer, editor, *Artificial Intelligence in Theory and Practice III*, pages 165–173. Springer, 2010.

[12] K. F. Yeung and Y. Yang. A Proactive Personalized Mobile News Recommendation System. In *2010 Developments in E-systems Engineering*, pages 207–212. IEEE, Sept. 2010.