

Reducing the Sparsity of Contextual Information for Recommender Systems

Dusan Zelenik
Faculty of Informatics and Information
Technologies
Slovak University of Technology
Bratislava, Slovakia
zelenik@fiit.stuba.sk

Maria Bielikova
Faculty of Informatics and Information
Technologies
Slovak University of Technology
Bratislava, Slovakia
bielik@fiit.stuba.sk

ABSTRACT

Our work focuses on the improvement of the accuracy of context-aware recommender systems. Contextual information showed to be promising factor in recommender systems. However, pure context-based recommender systems can not outperform other approaches mainly due to high sparsity of contextual information. We propose an idea to improve accuracy of context based recommender systems by context inference. Context inference is based on effect discovered by analyses of the context as a factor in uencing user needs. Analyses of the news readers reveals existence of behavioural correlation which is the main pillar of proposed context inference. Method for context inference is based on collaborative filtering and clustering of web usage (as a non-discretizing alternative to association rules mining).

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: Clustering Information Filtering; H.2 [Database Applications]: Data mining

General Terms

Algorithms

Keywords

context, recommender system, clustering, user behaviour

1. MOTIVATION

Context-aware recommender systems have become very popular since variety of contextual information could be acquired. With an increase of the smart-phone popularity and available features which they provide, we are able to associate user needs with contextual information. From the high level context types such as location, time, weather, to the low level context types such as humidity, noise, movement, we study the impact of the context on the user behaviour

and needs. However context itself has shown to be insufficient when it comes to accuracy of context-aware recommender systems. Context is therefore used as a secondary aspect for generating recommendation.

One reason for low accuracy is high sparsity of contextual information. High sparsity is caused by various natures of users and their preferences [5]. Some users do not want to share their personal information such as location, thus causing missing contextual information. Poor context information leads to low accuracy in prediction. On the other hand, some users are willing to expose even personal contextual information such as emotions. They are willing to answer question and explicitly express contextual information, which is then useful in context-aware recommendation.

Our idea is to propagate contextual information from one user to another in order to reduce the sparsity of data. We propose the propagation of the context by exploiting a correlation in users' behaviour. We assume that users' behaviour is not random, it is based on context of the user. For instance, Perse discovered association between negative mood and tendency to watch competition-style programs as a result of the need to experience happiness. Action-style programs are selected when viewers are in a positive emotional state. However, even if some associations are valid for majority of users, we expect that there are associations which could be discovered only for a subset of users. This leads to clustering of users by their behaviour. Identifying clusters of similar users helps to identify how to propagate the context between users.

We discover associations between the need and context using alternative to standard association rules mining. The difference is in non-discrete values which we use. For example, wrong discretization causes noise, as we lose the ability to compare them and thus sort them. Therefore we expect to achieve higher accuracy when we use proposed value based associations discovery instead of item based. To accomplish value based associations discovery we combine standard techniques from machine learning such as x-means [6] and vector distance computation (Euclidean distance).

2. RELATED WORK

Context inference has received relatively little attention in the literature when it comes to implicit inference from the logs of user activity. It is caused by the selective approaches

to context incorporation. Specific solutions work with specific contextual information. Kahng et al. demonstrate the predefined context as one of the factors for document ranking in information retrieval process. As an example they introduce weather and its impact on the user's interest in song listening. This empiric context selection emerges from the observation made by Baltranas et al. [2]. They research the relevance of the context in the system explicitly by asking the user. They showed that supposed context has positive impact on the success of their method. On the other hand, research by Asoh et al. [1] proves that there is a significant difference between the real and supposed reaction to the context. One way or another, this could be understood as an explicit form of the context acquisition. And unfortunately, we are still unable to persuade and engage everyone into explicit feedback.

Therefore we work with the acquired context and users' behaviour to infer missing contextual information. To stress the unavailability of contextual information we pick the work of Bermingham et al. [3]. They propose a solution to discover the sentiment from microblogs. The sentiment is a derivation of the emotional context. Microblogs are perfect source for discovering this type of the context. However it is domain specific and could not be used as a generic solution.

Riboni et al. announced a hybrid of statistical analyses and ontological reasoning in order to acquire the context. Utilization in the COSAR project shows better results by combining both of these approaches. We have decided to use statistical approach boosted by empirically observed effect of users behaviour correlation. We understand the correlation of the behaviour as the correlation of the contextual information. Konomi et al. present connections between people formed by co-presence at places. These connections are based on geo-location but correlate with social connections.

User behaviour is often represented by a set of actions performed by the user. Kramar observed the effect of changing the behaviour with the change of current context. He identified that multiple personas are present in the behaviour of individual. The same effect was exploited by Park et al. . They clustered user's behaviour by actions to improve query suggestions. They have actually used client side logs to cluster the behaviour, which outperformed the state-of-art approaches. From multiple personas of an individual, we expanded to multiple personas of all users in the system. Our intention is to supplement missing contextual information using multiple personas.

Combining multiple personas of more users will improve results in context inference. Research made by Cadiz et al. [4] or Rahnama et al. enables us to work with more users and in various systems. By using standardized frameworks and unifying context-aware systems, we are able to gather usage logs. Contextual information on activities from various systems improves our abilities to infer missing context. The only drawback of such framework is the redundancy of some information and higher complexity. Including information on the past, present and future context increases the complexity even more. Several approaches have been presented

to address this problem. Komninos et al. work with vector representation of action and propose solution to reduce complexity, even the complexity caused by vector weighting issues. Reduction of the complexity is important even for mobile devices where computational resources are constrained. Dargie et al. discuss the need to reduce time to recognise the context and its essence in real-time systems.

3. CORRELATION IN SIMILAR USERS' BEHAVIOUR

We have studied the effect of correlation in users' behaviour to propose an exploitation which would help to reduce the sparsity of contextual information. Our idea is to propagate contextual information only to users whose behaviour highly correlates.

3.1 Contextual Information

To prove our concept we have decided to work with database of web usage recorded by news portal SME.sk¹. This news portal is the biggest local news portal with more than 20 thousands active readers at the peak. Every click recorded includes time, IP address, user identifier and article identifier. Further information such as category, section, author, publishing time for article are also provided. We used this database before for content-based recommendation which enables us to compare results achieved by our previous work. To prepare database for further research, we add the context which is not in database. We use services such as wunderground² and ip2location³ to add information on weather and location. We also process timestamp to store time derivatives (such as day of week, part of day, etc.). Location which is extracted from IP address is very rough and for dynamic block of IP addresses, the location is almost untraceable. We have also applied a simple rule based algorithm to extract information on location (home, work, outside) using time and IP address. It is based on repeating IP addresses during work days (from 8 AM to 5 PM), during night and weekends. If the IP address was used by user during work hours many times, we add the context of location respectively (at work). Dataset prepared in this way contains contextual information which is acquired with both high and low confidence. Low confidence causes sparsity what negatively affects further recommendation process.

3.2 User habits

We have analysed news reading with focus on various context types. We presume that user has habits which are affected by context. We also presume that some users have similar habits thus their behaviour is affected by context similarly. We have mostly analysed the time as the most popular context (see Fig. 1). The figure shows that majority of users are influenced by forthcoming events. The figure proves that majority of users have similar habits. For instance, they read about cooking, when they are going to cook for Christmas. We also recognized same habits in smaller groups of users. For example, local football games are commented on this site, which attracts some users with interest in football. These users have same interests. It could also

¹<http://www.sme.sk>

²<http://www.wunderground.com>

³<http://www.ip2location.com>

mean that these people are also similarly influenced by the same context. Influence of context means correlation in their behaviour. We use this effect of correlation in behaviour to form clusters of similar users. Knowing similar users enables us to propagate contextual information correctly.

Percentual share of displayed articles during a year (shares for three different categories)

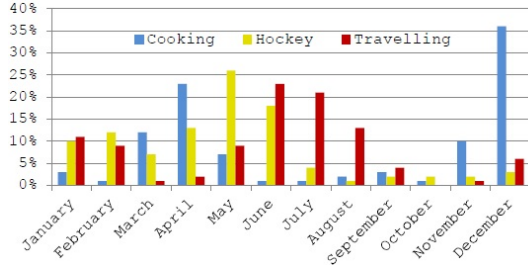


Figure 1: Readers are influenced by forthcoming events. We can observe the impact as an increase of displayed articles on specific topics (monitoring real users at SME.sk).

4. USER MODEL ENRICHMENT

Our user model represents the measure of user interest in item. Similarly to association rules we work with patterns. Every pattern consists of a condition and filter. In our case of news recommending, the filter is a combination of the section and the category in news portal. There are around 420 combinations which could be used. Condition expresses the context of the user which has to be valid when the rule is applied to recommendation process. We understand condition as a set of contexts which form a condition together. Conditions are used to find situation of the user. Condition and current situation of the user must be matching when we want to apply the filter. User model contains only the most frequent patterns. But even if condition does not match current situation of the user, we are still able to find the best matching condition. Every context has its value which represents the importance with a condition. Calculation of vector distances between conditions and situation of the user results in best matching rule which is then applied.

4.1 Context Inference

We have built the user model by processing user activity which has been already recorded. Some contextual information could be acquired directly using services and processing attributes. However, some contextual information could be missing or the confidence of the information is very low. We propose the context inference which leads to the reduction of the sparsity. Table (see Tab. 1) demonstrates how could the contextual information be propagated to another. These actions could be represented as vectors and clustered using all attributes except missing location. User A has complete contextual information in this example. User B is considered to be similar to User A since their behaviour is similar. In result, contextual information for missing values is inferred using known values and similar actions.

Table 1: Example of context inference. Actions 1, 2 are similar to actions 6, 7. Unknown context of location could be propagated to user B using known location from user A.

action	user	hour in day	day in week	section	location
1	A	22	2	sport	work
2	A	21	3	sport	work
3	A	11	7	cooking	home
4	A	11	7	cooking	home
5	A	12	7	cooking	home
6	B	19	3	sport	N/A
7	B	18	5	sport	N/A
8	B	12	7	cooking	N/A
9	B	12	7	cooking	N/A
10	B	13	7	cooking	N/A

1. **Cluster users' actions.** User identifier is ignored at this moment. We use x-means clustering [6].
2. **Compute users' similarity.** Co-presence of actions made by different users reveals their behavioural similarity.

$$similarity(u, v) = \frac{\sum_{i=0}^n (|U \cap C_i| + |V \cap C_i|)}{|U| + |V|}$$

where U and V are sets of actions for users u and v and C is the set of clusters.

3. **Find the best matching actions.** Only actions of similar users are used.
4. **Propagate contextual information.** Only in case it is available.

Context inference is followed by enriching the user model. It is done by finding the most frequent actions and extraction of conditions and associated items. In order to keep values not discretized we use clustering to find clusters of similar actions per user. Complexity of such computation is however very high, thus we process only the most recent actions. This also keeps the user model up-to-date.

4.2 Recommendation

There are more options how to incorporate context into recommender systems. Our user model enables us to recommend items by filtering them using rules stored in the model. Current situation of the user is used to find the best matching items. Items are used as filters on the dataset of potential recommendations. Here we can work with content-based approaches, collaborative filtering or other. Content-based recommendation is generated using items which are fetched from the model using current user conditions. In this alternative we are searching for the items in the dataset which are similar to items from user model. Item in the user model is not necessarily one of items in dataset. It could be only the set of keywords. We propose to use category and section in our dataset of news. Collaborative filtering is another very popular approach to generate recommendations.

To use our model with this approaches we change the pair of condition and item to condition and user. This enables us to reveal the most similar users whose situations were very similar to the current situation of the user.

5. EVALUATION STRATEGY

As we already mentioned we work with database where both low and high confident contextual information is present. We propose to evaluate our method for sparsity reduction by using only records with high confident contextual information. We randomly select contextual information and simulate its confidence to be very low. Then we apply context inference and compare inferred results with original values. We also propose evaluation for context-aware recommendation. We want to compare results achieved by recommending news both with and without inferred contextual information. Experiment is conducted with real people who are separated into two groups. One group receives recommendations generated only with original contextual information. Another group receives recommendations generated with inferred context (reduced contextual sparsity). In this case we also incorporate test for statistical significance. Another approach to evaluate our approach to sparsity reduction is to compare our recommender system to others to show expected improvement.

6. CONCLUSIONS AND FUTURE WORK

In our work we face significant drawback of common sparsity in contextual information. We presented our proposal for solving sparsity in contextual information using context inference. We showed analyses of the web usage for news portal and revealed the effect of behavioural correlation. Clustering users by the web usage splits users with similar preferences into groups where the association between context and need is also similar. In such group we can propagate missing contextual information or context value which is lacking in confidence. Our approach solves problems which are often present in frameworks which are gathering contextual information from more sources. In our future work we plan to generate news recommendations using inferred context and compare results to our previous work where we used pure content-based recommendations. We also plan to apply this method for smart-phones where we encounter higher variety of context types. Context which could be acquired on smart-phone is more complex than on the web what is big challenge for us. Our main contribution is in reducing the sparsity of contextual information thus improving accuracy of context-aware recommender systems. We have also designed an alternative to association rules mining which respects numeric values.

7. ACKNOWLEDGMENTS

This work was partially supported by the Scientific Grant Agency of the Ministry of Education of Slovak Republic, grant VG1/0675/11 and by the Slovak Research and Development Agency under the contract No. APVV-0208-10.

8. REFERENCES

- [1] H. Asoh, Y. Motomura, and C. Ono. An analysis of differences between preferences in real and supposed contexts. In *Proceedings of the 2010 workshop on context-aware recommender systems*, 2010.
- [2] L. Baltrunas, B. Ludwig, and F. Ricci. Context relevance assessment for recommender systems. In *Proceedings of the 16th international conference on Intelligent user interfaces*, pages 287–290. ACM, 2011.
- [3] A. Bermingham and A. F. Smeaton. Classifying sentiment in microblogs: is brevity an advantage? In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1833–1836. ACM, 2010.
- [4] A. Cádiz, S. González, and K. Mens. Orchestrating context-aware systems: a design perspective. In *Proceedings of the first international workshop on Context-aware software technology and applications*, pages 5–8. ACM, 2009.
- [5] D. Carlson and A. Schrader. A wide-area context-awareness approach for android. In *Proceedings of the 13th International Conference on Information Integration and Web-based Applications and Services*, page 383. ACM, 2011.
- [6] D. Pelleg, A. W. Moore, et al. X-means: Extending k-means with efficient estimation of the number of clusters. In *ICML*, volume 1, 2000.