MODELING THE DYNAMICS OF RESEARCH INTERESTS

Tomáš KRAMÁR, Roman BILEVIC

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
kramar@fiit.stuba.sk, roman.bilevic@gmail.com

Abstract. This paper presents an approach to modeling user's research interests and their change in time. Human interests are not stable and modeling approaches should take this information into consideration. We present a model that captures the influence of time and uses keywords to represent the user interests. The algorithm that is used to create model itself is known as divisive hierarchical clustering (DHC). The information about changes in user interests are captured when the user is browsing on the internet in form of implicit feedback. The modified three-descriptor representation is used to express the time influence on user interests.

1. Introduction

This paper presents an approach to collecting contextual information about the user and its usage to improve the search engines. The contextual information is represented by the user interests in form of a lightweight, keywords-based semantics. The proposed method considers influence of time onto the user interests and evaluates actual interests as the most important source of contextual information.

In this paper we understand the context as information specific to one user (e.g. hobbies, job description, GPS location etc.). This information should constrain the field of knowledge in which user wants to search for information. We can also say that user has multiple contexts if in some situations we want to search for distinct information in the same field of knowledge, for example: we want to find solution for some programming problem, but at work we use different programming language than at home and therefore we have two contexts of location. When we write about word context, it means that the word is ambiguous and it is connected with multiple fields of knowledge.

Human interests are not stable, they change over time. People can pick up new interests or lose the old ones. Actual priority of the interests can change several times a day. From temporal perspective there are two categories of interests: long-term and short-term.

Long-term interests [1] can last for many years: short-term interests can last from couple of hours to several months. The short-term interests have more impact on immediate user needs.

The dynamics of the interests bring two main issues. The first one is called *"burst"* [4]. In the moment when users gain new interest, they tend to consume much information related to that interest in a short period of time. It is because they want to get at least a basic knowledge about that interest. In such situation, the user model should capture the need to obtain information about new topic. The second issue is called *"drifting"* [6]. It is the situation when user changes interest several times during the short period of time.

2. Related work

There are few approaches that consider influence of the time on user interests. Some of them use time window to determine the most recently obtained information that will be considered in modeling of the user context. This approach was developed to solve issue with "drifting". Partial memory method presents forgetting mechanism [7] that gives information context weight, which determines the extent of its impact on the search results. Information in time window gets full context weight. The older the information outside of this window is the lower context weight it gets. Another method using time window is hybrid user model [3]. It primary uses time window to determine the user context. If the information in time window is not sufficient to determine user context then all information is used to do so.

Three-descriptor was developed to explicitly capture the change of user interests in time [10]. It was assigned to every document in a digital library. After opening and reading a document, user had to give explicit feedback whether he is actually interested in the document or not. Three-descriptor consists of positive and negative short-term descriptor and long-term descriptor. One of the short-term descriptors is increased after every feedback and the other one is decreased. If the user was interested in the document then the positive one is increased and the negative one is decreased. Their values are from interval <0, 1>. Long-term descriptor is not modified by user's direct feedback. It is calculated after every feedback using short-term descriptors:

$$LTD = LTD + \beta\, Dpos - (1 - \beta)\, Dneg \quad (1)$$

Long-term descriptor is LTD, β is weight coefficient (usually equal 0.5), Dpos is value of positive short-term descriptor and Dneg is value of negative short-term descriptor.

3. Temporal dynamics in user model

We propose a process to introduce temporal dynamics into a user model. This process consists of five partial stages, which are described separately in the following text. Information about the user is gathered using keyword extraction. These keywords are used to create the model of user's interests. Interest's actuality is computed using the three-descriptor which captures the time information.

Keyword extraction. The user interests are represented by keywords [2]. They are extracted from every website that the user browsed, so we can get complex information about user's interests. It is also important how the user got to the website: by typing the

URL address to the location bar of the browser, clicking on the link on a generic Web site or clicking on a search result in a search results page. Clicking on a searched result means that user search for information purposefully, therefore sites that were visited through the search results tell us about user interests more than potentially random browsing.

Apart from extracted keywords we also record their mutual occurrence on the websites. Mutual occurrence indicates that the keywords are related to one topic and can therefore possibly represent user interest. The higher the mutual occurrence of keywords the stronger is their contextual relation.

We also record the time the user has spent browsing the site. If the time between opening and closing the website is less than a predefined threshold, it means that user has quickly evaluated the information on the website as not interesting. It gives us implicit feedback about his actual interests.

Model of interests. We use the divisive hierarchical clustering algorithm (DHC) [5] to create a model from all the extracted keywords. The nodes of the model are keywords and the strength of a connection between nodes is equal to the number of mutual occurrences of keywords the in browsed websites. DHC is a recursive algorithm that divides cluster of connected keywords into smaller child clusters. First cluster contains all the keywords. There are four steps that have to be done in every cycle of recursion:

1. check if cluster can be divided
2. determine the connection strength threshold
3. remove connections that are weaker than threshold
4. create new clusters from nodes that are still connected

Cluster can't be further divided if it meets one of the following conditions:

- number of nodes in cluster is less or equal to the minimal count
- there were no connections removed in the previous cycle of iteration, which means we can't find more weak connections to delete

Connection strength threshold can be determined explicitly or dynamically. We determine the threshold dynamically using the *"Valley"* algorithm [8]. This algorithm is suitable for large data sets. This algorithm creates a histogram from connection strengths in a cluster. The histogram is then divided into two parts. The split point has low density and is from both sides surrounded by points with high density. If there are more than one such split points, the steepest one is used. This point is then denoted as connection strength threshold.

After removing all the weak connections we copy the nodes that are still connected together to the new child clusters. The original DHC algorithm copies every child node only once. Since there may be ambiguous keywords that can have more than one context, we modified DHC algorithm. Node in one child cluster can be copied to another child cluster if it had threshold strength connection with any of the clusters nodes.

If no cluster can be further divided, final clusters are denoted as leaves of the model. Keywords in one leaf of model express the nature of one specific interest and leaves of created model should match the user interests.

Representation of the time influence. The original idea of three-descriptor [10] was to rate the actual user interest in documents based on explicit feedback. We modified it to rate whole interests based on implicit feedback.

The influence of time on user interests is represented by three-descriptor that is assigned to keywords connection. The connection expresses the relationship between keywords and carries the context information. One ambiguous keyword can occur in multiple leaves of the model representing multiple interests. Each interest behaves differently in time and therefore each meaning of the keyword should have separate three-descriptor. Therefore it is better to assign three-descriptor to the connection than to the keyword itself. The basic idea of components of three-descriptor is preserved but the way their value change was reworked to meet the need of implicit feedback usage. The long-term descriptor shows how long user has certain interest. Short-term descriptors show how actual that interest is. After every keywords extraction, three-descriptors of extracted pairs of keywords are adjusted.

Positive short-term interest descriptor increases when user dwells on the website longer than predefined time limit. If user got to the website through search results list, the descriptor increases even more.

Negative short-term interest descriptor increases when user dwells on the website shorter than predefined time limit. The descriptor increases even more if user clicked on an item in a search results list but omitted top results. The user just by looking at the results list concluded that first results don't contain desired information.

Long-term interest descriptor doesn't increase after every extraction. It increases mostly once a day to preserve the real length of the interest.

Time limit is a threshold value that represents the minimal time user has to spend reading the website to show interest in its information. It is computed dynamically for every visited website based on its text length and readability ease. Average reading speed is 250 words per minute [9]. To get an understanding what information website or document contains user usually doesn't read the whole text. In scientific documents user usually reads abstract of average length 100-150 words. In unstructured texts users tend to quickly scan the text reading only every tenth to fifteenth word. Combining these numbers we get that reader needs maximum of 24 seconds to evaluate if website contains the information he is interested in. This applies for the text that is at least 1000 words long. For shorter texts is this time reduced proportionally. We also use Flesch readability index to scale this time appropriately to reading ease of the text.

Interest evaluation. We use components of three-descriptor to evaluate actual priority of the user interests. Partial actuality is computed using following formula:

$$a = \log\left(\frac{\mathrm{psd}}{1+\mathrm{nsd}}\right) * \mathrm{ld}^{-1} \qquad (2)$$

Actuality is *a*, *psd* is value of the positive short-term descriptor, *nsd* is value of the negative short-term descriptor and *ld* is value of the long-term descriptor.

New interests have high *psd* and low *ld* and *nsd.* This helps to deal with *"bursts"* which is one issue of the temporal dynamics. During *"burst"* is *ld* usually equal to 1, so even relatively low values of *psd* result into high actuality.

Since there is more than one connection in every leaf of model, the actuality of the whole interest is computed as an average of all partial actualities in the leaf.

4. Experimental evaluation

In order to validate our approach to modeling the user interests, we conducted two experiments.

The goal of the first experiment was to evaluate the hierarchical user model based on the lightweight semantics. We have collected data from one user about browsing on the open Web. We have used the described approach to create the hierarchical user model and asked the subject to evaluate the final user model with respect to consistency at the leaf-level and general coverage of the interests. The user model as generated by the proposed method was as follows:

- Interest index 0.29: Jean-Philippe Lang, Redmine Sign, Redmine
- Interest index 1.33: CONSEILS DE DEGUSTATION, Bordeaux A.O.C, grande fraîcheur aromatique
- Interest index 0.25: Annota, bookmarks, important bookmarks, easier search, Firefox
- Interest index 0.19: Recommend rake db, Diff, rake db
- Interest index 0.72: end end, method, instance, class

The subject verified that the user model fits the interests and that the leaf levels are consistent.

The second evaluation was aimed at validating the capturing of the time changes. We have created two user models for all users of the Annota research tool. The second model was created 12 days later than the first model. We have sent both user models to each Annota user and asked them to evaluate if the difference in the models corresponds to the changes in interests during that 12 days period. From the total of 12 people who answered the questions, 11 confirmed that the identified leaf-level interests match their research interests. 7 people confirmed that the order of interests matches their strength. 3 people confirmed that the changes between the two models correspond to real changes in user interests, while 7 people did not observe any changes in the models and 2 people said that the changes were not captured correctly.

5. Conclusions

We have proposed a method to model changes in user interests. It takes into consideration the time influence on these interests. The keywords extracted from every website that user browse are used to represent the user's interests. The divisive hierarchical clustering algorithm is used to create model of interests. Keywords are used as nodes of the model. Since we cluster the keywords based on their meaning we had to modify DHC algorithm to copy ambiguous keywords to multiple clusters. Every leaf of the model contains the keywords that represent one particular user's interest.

The influence of the time is captured by three-descriptor assigned to keyword connection. With its use we can estimate how actual the user's interests are. We can quickly identify new interests and therefore we can solve the problem with *"bursts"* The three-

descriptor was modified to use implicit feedback in form of time that user dwelled on the website. The time limit is computed for every website based on its text length, average reading speed and reading ease. The user has to dwell on the website longer than the time limit to show the interest in the information it proposes. Also the websites visited via search results page have higher impact on actual user interests. Dealing with the *"drifting"* problem is best left to the system that uses the model. E.g., in case of a search engine, this problem can be solved by mixing multiple search results corresponding to each user interest.

The preliminary experimental results with the approach show that the directions are promising and that we are able to identify the changes in interests over time.

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

References

[1] Barla M., Tvarožek M., and Bieliková M. Rule-based User Characteristics Acquisition from Logs with Semantics for Personalized Web-based Systems. In *Computing and Informatics*, Vol. 28, No. 4, 2009.

[2] Bieliková M., Barla M., and Šimko M. Lightweight Semantics for the "Wild Web". Keynote. In *WWW/Internet 2011, Proc. of the IADIS Int. Conf. IADIS Press*, pp. xxv-xxxii.

[3] Billsus, D., Pazzani, M., J.: A Hybrid User Model for News Classification. In: *Proceedings of the Seventh International Conference on User Modeling (UM '99)*. New York: Springer-Verlag, (1999), pp. 99-108.

[4] He, D., Parker, D.: Topic Dynamic : An Alternative Model of 'Bursts' in Streams of Topics Categories and Subject Descriptors. In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD'10). New York: ACM, (2010), pp. 443-452.

[5] Kim, H., Chan, P.: Learning Implicit User Interest Hierarchy for Context in Personalization. In: *Proceedings of the 8th international conference on Intelligent user interfaces (IUI'03)*. New York: ACM, (2003), pp. 101-108.

[6] Koychev, I., Schwab, I.: Adaptation to Drifting User's Interests. In: *Proceedings of ECML2000 Workshop: Machine Learning in New Information Age*. 2000, s. 84-91

[7] Maloof, M., A., Michalski, R., S.: Selecting Examples for Partial Memory Learning. In: *Machine Learning*, (2000), vol. 41, no. 1, pp. 27-52

[8] Milenova, B. L., Campos, M. M.: O-Cluster: Scalable Clustering of Large High Dimensional Data Sets. In: *Proceedings from the IEEE International Conference on Data Mining (ICDM'02)*, (2002), pp. 290-306

[9] White, R., W., Bennett, P., N., Dumais, S., T.: Predicting short-term interests using activity-based search context. In*: Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM '10).* New York: ACM, (2010), pp. 1009-1018.

[10] Widyantoro, D., H., Ioerger, T., R., Yen, J.: Learning User Interest Dynamics with a Three-Descriptor Representation. In: *Journal of the American Society for Information Science*, (2001), vol. 52 , no. 3, pp. 212-225.