Grouping Web Users based on Query Log

Julia Kiseleva

University of Saint Petersburg
julianakiseleva@gmail.com

Abstract. Grouping web users is one of the most important research topics in web usage mining. Existing approaches grouping web users based on the snapshots of web user sessions. Web user groups generated based on their historical web sessions are useful in intelligent web advertisement and web caching. In this paper we focus on the grouping web users using their web logs.

Keywords: data mining, Web Usage Mining, WWW, Web logs

1 Introduction

Web usage Mining (WUM) - the application of data mining techniques to discover usage patterns from the web data has been an active area of research and commercialization [1]. Existing web usage data mining techniques include statistical analysis[1], sequential patterns[2] , association rules[3], classification[4]etc.

An important topic in web usage mining is grouping web users – discovering group of users that exhibit similar information needs. Group based on the user's query log and consolidate users with similar interests (we assume that interest is main part of query, e.g. from query "mountain bike" we suppose that user have strong interests for bikes). Further we define such set of queries as thematic slice of data. There are 3 main approaches to collect data for analysis of user behavior: search engine logs [7]; collection on client; proxy logs. Clearly, at client side it is possible to collect richest set of information about user behavior but amount of data is limited (because it is just 1 user), collection on server side provides access to a lot of data but depth of knowledge is limited. For our research we were using proxy logs. This way we can benefit from both having reasonably detailed user histories and working with set of histories of reasonable size. Web pages are personalized based on the interests of an individual. Personalization implies that the changes are based on implicit data, such as items purchased or pages viewed. In our research we don't approach to strongly user's personalization. Generally, typical web user grouping approach consists of three phases: data preparation, group discovery and group analysis. In the first phase, web sessions of users are extracted from the web server log by using some user identification and session identification techniques. Existing web user grouping methods based on the snapshots of their web sessions. However, the web usage data is dynamic in nature. Such dynamic nature of web usage data gives two opportunities to

web user grouping. Discovery of novel web user groups we can discover groups of users that exhibit similar characteristics in the evolution of their usage data:

- *Maintenance of web user grouping results*: Web user groups generated by existing techniques at time T1 does not include the usage data at time T2 and beyond. Hence, the grouping results have to be updated constantly along with the change of web usage data. This requires development of efficient incremental web user grouping techniques. With help to find user's groups we introduced similarity metrics for measure the likeness between web user's queries. Similarity is our criterion to define how close user to each other. This problem is open and hot at present time. Purpose of this research is to develop unsupervised algorithm to identify groups of users sharing same interests based on their requests to search engines. We assume that users may have multiple interests and interests are not known in advance. In particular we are interesting to find groups with follows properties:

- Not tiny (need large joint user histories for further analysis with machine learning);

- Not huge (want to be able to learn rules specific to relatively small subgroups of users);

- Group expressed thematic interest (like in our example about bike);

- Group which further we can refer to main class (or cluster) (like thematic interest "Vegas" refer to the main class "Travel" in our proposition).

This is work in progress report. At this stage of our research we focus on user similarity metrics that later will be user to group users. In this report we present description of our approach, define several metrics and conduct experiments to evaluate their quality.

2 Data Set

For our research we used sample log of user queries to 4 major public search engines based on proxy approach (log was provided by Nebuad (www.nebuad.com). The log contains 66380 queries asked by 1343 users over a month. Data set consists of unordered pairs (User Id, User Query) see sample below in Table 1:

Table 1. Table contains example for data.

10007_6	Sun
10007_6	botanica
10007_6	man
10007_6	Sun
10008_6	booty
10008_6	Sun rise
10008_6	burning man pictures
10008_6	booty

2.1 Data cleaning

Before beginning our experiments we "clean" user's queries using two methods:
1) remove stop words like "www", "com" and etc. and punctuation marks;
2) remove words which don't appear in WordNet [6].
Cleaning reduce user's dictionary (set of using words) on 1.4 %. After cleaning data don't contain stop words.

3 Approach

There are many different approaches to group data such as clustering, etc. However, any grouping method assumes that there is metric to measure similarity between users. Therefore, as first stage of this research we decided to investigate possible definitions of such metrics and evaluate their usefulness. They are not new but we need to understand what disadvantages they have with purpose to improve them. And as result get unified user evolutions methodic.

3.1 Mapping queries to vectors

To define metrics we need to represent queries in the vector space first. General approach is the same for all our metrics:

1 For each user we get a vector of terms that appeared in the document:

$$d_j = w(t_1), w(t_2) \ldots w(t_n)_j. \quad (1)$$

where $w(t_j)$ is the tf–idf weight of term t_j (term - frequency–inverse user's query frequency) [5], where

$$\mathrm{tf_{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}}. \quad (2)$$

where $n_{i,j}$ is the number of occurrences of the considered term in document d_j, and the denominator is the number of occurrences of all terms in document d_j.

The inverse document frequency is a measure of the general importance of the term:

$$\mathrm{idf_i} = \log \frac{|D|}{|\{d_j : t_i \in d_j\}|}. \quad (3)$$

- Where $|D|$: total number of documents in the corpus.

- $|\{d_j : t_i \in d_j\}|$: number of documents where the term t_i appears.

2. Each user was represented as a set of vectors of queries that appeared in the document giving an answer to the query:

$$Q_i = w(t_1), w(t_2) \ldots w(t_n)_j. \quad (4)$$

where $w(t_j)$ is the associated weight of term j inside query Q_i.

We try to find the most optimal evolution similarity between users, to achieve our purpose we realize experiments for the methods above. As a result we have matrix of closeness between all users, where $\{a_{ij}\}$ is measure of closeness between user *i* and user *j*.

3.2 Metric 1: Averaged metric user

For this metric we model user as union of words from his queries. E.g. for user 10008_6 from example table his model is {booty: 2, sun: 2, pictures: 1, man: 1, burning: 1} applying TFIDF procedure from previous section to such sets (using them as documents). Obviously, most similar vectors will have maximum quantity of scalar product. According to this metric user is most similar to itself. If there are no common words then similarity is 0. If two users share rare word then it is better than sharing popular word.

3.3 Metric 2: Maximum query similarity

For this metric, we define closeness between two users as maximum value scalar product of their queries vectors of terms weight. E.g. two users have equals queries, closeness between them =1. Such metrics is not sensitive for cases when user U_1, user U_2 and user U_3 have similar queries (e.g. “vegas”), closeness between them is 1, but we can meet situation like U_1 entered query “vegas” 10 times, U_2 - 8 times, U_3 - only 1 time. In such cases our metrics works not so well.

4 Preliminary Results

At present time we at the begging of our research and have only preliminary results. One major complication is difficult to evaluate quality of observed results. We describe our approach to evaluation in next section.

4.1 Evaluation

At present time we at the begging of our research and have only preliminary results. Evaluation problem is as follows: given 2 alternative results for user U 1 we need to pick which one is best. Result is set of pairs (user_id, score). This can be considered as orde- red list such pairs without any loss of generality. One metric is calculated how many users are considered to be similar to user u. Also it does not give an idea is it good or bad. However, together with expert answers on whether these pairs of users are actually similar this helps to get some quantities measure. To collect expert answers we use "topic slices" or "word slices" of data. Starting with particular "good" word that is picked manually we select all users that use this word and manually check if human expert thinks some other are actually similar given their query logs. Then we compare this to what our automated procedure returns. Examples of these words are "bike", "Vegas" and etc.

Yet another idea is based on observation that even if similarity between 2 users is not zero it does not mean they share any interests. E.g. this may happen because they used same popular word in one of their queries.

However, it is not obvious how to pick threshold to separate those who similar from those who are not. Our approach is following: we are going remark on the graphic of score's distribution with purpose to find quickly drop on it and using it as a threshold to separate.

4.2 Results

First experiments we did without filter for words using all words from log. Using WordNet as a filter for our logs we remove stop words like "jurt" they seem like user's error. Such action improved quality of getting groups. Also we remark that using WordNet helps us to group users, e.g. user U_1 have query "online games", user U_2 "games" our filter remove "online" and we'll get good similarity between users. In the diagram for the first averaged metrics (x-axis - quantity of users, y-axis – value of similarity between users) we can see distribution's graphic of scalar product's values for the first metrics. In the diagram for maximum query similarity metric (x-axis - quantity of users, y – axis – value of similarity between users) we can see distribution's graphic of scalar product's values for the second metrics. One of ideas after analyze this graphics is to use a smoothing method.

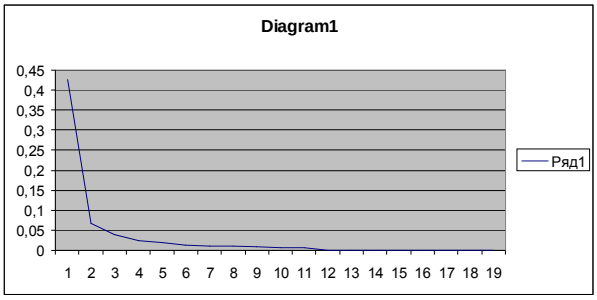

Fig.1 average metric similarity

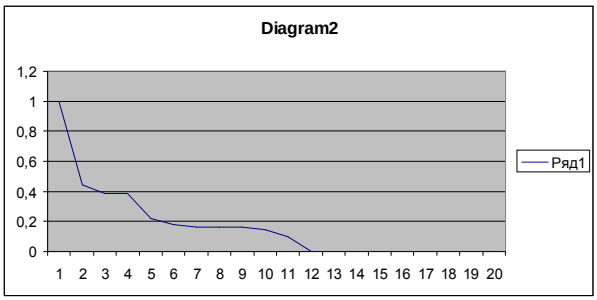

Fig. 2 maximum query similarity

5 Future Directions

In the future we plan to investigate other metrics, to analyze result as getting groups from Web user, moreover to know category initially of Web users is not a necessary condition. One way to do it is analysis thematic slice of data. We are going to get a map of Web users which based on user's interests. Using this map we can classify users by interests. We plan to extent input data with purpose to improve a quality of getting group, we are going add time.

6 Conclusion

The paper describes general ideas for grouping metric. These are the first steps, which have some effect and must be developed to obtain further results. This research is realizing under supervision of Boris Novikov.

References

[1] C. Buchwalter, M.Ryan, and D. Martin. The state of online advertising: data covering 4 th Q 2000. In TR Adrelevance, 2001.

[2] Q.Yang, H.H. Zhang, and T.Li . Mining web logs for prediction models in www caching and prefetching. In *Proc.of ICCNMC'01*, 2001.

[3] B.Mobasher, H.Dai, T. Luo, and M.Nakagawa. Effective personalization based on association Rule discovery from the web usage data. In Proc. Of WIDM,2001.

[4] T.Li, Q.Yang, and K.Wang. Classification pruning for web-request prediction. In Proctor WWW, 2001.

[5] Baez-Yates, R., Hurtado, C., Mendoza, M.: Query recommendation using query logs in search engines. In: Current Trends in Database Technology – EDBT Springer-Verlag GmbH (2004) 588– 596

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