Trending Words in Digital Library for Term Cloud-based Navigation

Samuel Molnár, Róbert Móro, Mária Bieliková
Institute of Informatics and Software Engineering, Faculty of Informatics and Information Technologies,
Slovak University of Technology, Ilkovičova 2, 842 16 Bratislava, Slovakia
{xmolnars1,maria.bielikova,robert.moro}@stuba.sk

***Abstract*—The clouds consisting of tags or keywords provide an alternative and often more readable navigation interface by exploiting visual features of words placed in a cloud and augmenting their information value with different font size and color. However, existing approaches for cloud navigation rely mostly on frequency of terms and do not adapt to the users' needs. In the paper, we propose a method for term cloud navigation which exploits navigation history as a source of metadata for personalized navigation. We consider trending words in users' navigation history as a relevant factor determining users' interests while navigating. In addition, we recognize a position of a word in a query to have an important role and rank the list of the documents accordingly. We focus on the domain of digital libraries and provide an evaluation of our method in Annota[1], bookmarking and annotation web-based system.**

Keywords*—*tags; keywords; navigation; history; term cloud; digital libraries

I. Introduction

Nowadays, the amount of information available on the web is making navigation by many common approaches and technologies difficult which leaves users to rely solely on the results list provided by the keyword-based search engines. However, search engines were not intended primarily for navigation and can thus provide only a little support for users in situations when their information need is more general or ill-defined.

Therefore, over the past years several novel approaches were presented. Some of them aim to enrich the traditional search with navigational support [9], or utilize the power of search engines to build recommendation systems that help users to find documents fitting their needs or interests [8]. Other approaches aim to provide new visual means of the information space exploration and sense-making, such as cluster-based views for search results browsing [6] or tag clouds [3], [1].

In this paper, we focus on the latter and introduce a method for cloud navigation which utilizes navigation history as a source of metadata for personalized browsing of information. Tag clouds traditionally exploit different visual features of words like font size, color or justification to emphasize their relevance and thus aid a user's navigation with the knowledge of how large the information space behind the specific word is or how the word is relevant to a user's current context. They also provide users a convenient way to refine their queries and discover new topics that are similar to their information need. However, the conventional tag cloud navigation does not reflect the needs and interests of the users.

In our proposed approach we consider trending words in the users' histories to be the closest expression of their interests in the explored time period. In addition, we take into account the positions of the words in the query (i.e. a sequence of terms selected by a user from the cloud) and modify the ranking of the resultant documents accordingly.

We evaluate our approach in the domain of digital libraries and specifically on the scenario of a researcher novice, who explores the domain in order to acquire the overview of the topics and research trends as well as to find relevant research articles.

II. Related Work

Enhancement of navigation within information domain is discussed in several research works that aim particularly towards utilization of visual features of tag clouds to improve navigability in the domain as well as selecting what document attributes should be used as the content of a tag cloud.

In [2], the authors propose an approach for text justification for creating more readable tag clouds. They use different typesetting algorithms, such as Knuth-Plass justification algorithm, combined with metrics for word similarities in order to create clusters of similar words in a cloud. In addition, the authors explore techniques for automated design scheme concluding that the alignment by nested tables is the most practical approach.

Gwizdka, et al. [1] introduced a novel method for tag cloud navigation by taking history into account. They use pivot browsing, so in each step of the navigation the content of the tag cloud is adapted to a current user's query. The history is visualized by three different tag clouds representing the three most recent steps in navigation. Tag clouds are distinguished by different colors and shared occurrence of tags (same tag in different clouds) is highlighted with the same color. By highlighting co-occurrences of tags authors demonstrate coherence in navigation and similarities between the words in the user's query.

Enhancing navigability using tag cloud and social data is presented in [4]. Authors propose a method for computing a set of relevant tags for a query specified by user. They define relevancy of a tag as a number of occurrences of a specific tag in documents retrieved for a user's query and a global number of occurrences of the tag. They enhanced the method

[1] http://annota.fiit.stuba.sk

by using social context such as preference of tags added by the user's friends. In order to evaluate their method authors conducted an experiment with students that were asked to navigate using Last.fm tags. The results of experiments were evaluated considering the time that it took a user to find a desired or familiar song from Last.fm.

Koutrika et. al. [3] present navigation and search over structured data using data clouds. Data clouds represent tag clouds that consist of precomputed search results over structured data. The purpose of the different content for the cloud is to guide users to refine their search query. Authors introduced a solution for a search refinement where entities span over multiple tables in database. A few approaches for scoring of keywords in a tag cloud are presented such as popularity, relevance or query dependency. The method for data clouds was evaluated using CourseRank, a system for ranking classes at universities and three students that were given a task to find a course in the system. The conclusion of the experiment was that the query dependency method for scoring keywords is the most efficient when refining search.

Existing approaches consider in most cases only tags as a source of metadata for content of the cloud [7] exposing their solution for cold start problem when recently added documents without any tags cannot be included in results of user's navigation. We try to remedy this problem by utilizing keywords automatically extracted from documents as source of metadata as well. Although Gwizdka, et al. [1] consider navigation history, it is only for visualisation of users queries in the short-term period of time without exploiting their navigation history. Therefore, their approach like most of the other presented lacks any personalization of the cloud content, thus providing only generic navigation in the specific domain.

III. History-based term cloud navigation

We propose a method for navigation which takes query history in a specific period of time into account. By using this approach we utilize the users' interests in that time period to enhance their navigation in the information space. We consider following history factors as relevant:

1) A position of a term in a query, which is represented as a sequence of terms selected by a user from the cloud
2) Trending terms in the users' history of queries in a specific period of time
3) Visualization of the term cloud in which terms from history are highlighted by different colors according to the time of their last usage.

Each navigation session starts with a tag cloud representing the whole information space or with an initial search specifying the part of it that is of interest to the user. However, we extend the cloud with relevant words from history utilizing history of all the users, as these words reflect the various paths foot-worn by the users seeking for information and can, therefore, help novice researchers find new topics to explore.

A. Term cloud content

In order to represent documents, we use tags created by users to describe and categorize documents as well as keywords extracted from documents which we both denote as terms. Since our method is mainly focused on documents within digital libraries' domain, we use the knowledge of their predefined structure to extract keywords only from relevant parts of documents, e.g. an abstract of an article.

When selecting terms into cloud, the relevancy of each term is determined by the number of times the term occurs in documents matching the user's query using the following formula:

$$R_w(w) = count(w, D) \tag{1}$$

where D is a list of all documents in domain that contain term w and $count$ is a function determining the number of occurrences of term w in the list of documents D. This is a basic generic relevancy measure and it could be further enhanced by using precomputed weights in the automatically acquisitioned lightweight domain model [11]; however, our focus is mainly on the cloud content personalization.

B. Position of a term in the query

The position of a term in a user's query during the navigation is used for ranking of the list of relevant documents containing at least one of the terms from the user's query. The relevancy of a term, as stated in (1), is adjusted by the term's position in the query by $log(n)+1$, where n is a position of a term, in order to prefer the newly selected terms and penalize the older ones. Thus, we prefer documents that are the most relevant to the current information need of the user based on an assumption that in each step of navigation the information need evolves as the user explores the domain. The following equation is proposed to compute relevancy of the document d:

$$R(d) = \frac{\sum_{w \in Q \cap W_d}(log(pos(w)) + 1)}{|Q|} \tag{2}$$

where Q is a query, $pos(w)$ determines a position of term w in Q and W_d is a list of terms (extracted keywords and tags) from document d.

C. History in a period of time

History records from a specific period of time are used for adapting the content of a cloud according to the user's query. When user's query changes, we search the history for similar queries and the most frequent terms in the queries are considered as the most relevant for the user's query. Thus, we help users to customize their queries with terms that they or other users have already used before.

This is especially useful for the scenario of exploratory navigation, because it helps users discover and explore what other users have been looking for. However, our approach can be easily modified for the refinding scenario as well. By utilizing history records of only one user instead of all the users, we can provide that particular user a way to quickly select already used navigation paths and refind potentially relevant information.

We apply the value of frequency by adjusting the relevancy formula (1):

$$R_w(w) = log(fq(w, p)) * count(w, D) \tag{3}$$

where $fq(w, p)$ is a frequency of term w in the period of time p. It is used in the process of choosing terms to cloud as described in Algorithm 1.

Algorithm 1: Choosing the words to cloud from history record and documents.

Input: Q - User's current query
W_d - Terms from relevant document for query Q.
MAX - Maximal number of terms from history.
Output: $Result$ - List of similar terms to the current user's query.

```
begin
    find queries in user's history that contain at least
    one term from query Q;
    choose up to MAX terms from found queries with
    the highest relevancies and assign them to W_h;
    foreach w in W_h do
        if w ∈ W_d then
            compute relevancy of w according to (3);
        else
            set relevancy of w to the mean relevancy of
            W_d.
        end
    end
    add W_h to Result;
    return Result.
end
```

D. Word cloud visualization

Our approach for representation of history in a cloud exploits the time of the last usage of the terms from cloud in users' queries by using different colors as shown in Fig. 1. A color of a word is computed by interpolating between two colors based on the time of the last usage within the explored time period as follows:

$$color(w, p) = interpolate(CLR1, CLR2, norm(w, p)) \quad (4)$$

where $color(w, p)$ is the color of the term w in explored time period p and $norm(w, p)$ maps last usage of term in explored time period on $\langle 0, 1 \rangle$ interval.

By employing visual color distinction for terms used in history, user is able to see how recently she or others navigated in similar navigation path to hers. This personalization step helps especially novice users to see queries of different users navigating with similar information goal.

IV. Evaluation

We implemented our method in Annota, which is a system for creating bookmarks and annotating documents [10] and it is being developed at the Slovak University of Technology in Bratislava as a part of a ongoing research project TraDiCe (Cognitive Traveling in Digital Space of the Web and Digital Libraries) [5]. Currently, it is used by more than 100 users having about 4700 research documents bookmarked. Apart from bookmarking, Annota allows the users to to collaborate by creating groups and search in bookmarks and articles marked as public.

Our method is implemented as a module into Annota (see Fig. 1). We use Elasticsearch[2] for indexing documents and history of users' queries. We use Elasticsearch facet search to compute relevancy of terms in documents matching user's query as described in (1). Keyword extraction is backed up with AlchemyAPI[3]. Navigation interface is implemented using Backbone.js JavaScript framework. We gather navigation data represented as a list of queries users constructed during navigation, number of words in a cloud added from documents and from history. Apart from these logs we also track which documents users visit during their navigation sessions.

We conducted three experiments to evaluate how the users navigate using our method when their information need is more general and the main goal is to explore the domain. The first experiment served as a comparison of users' queries during the navigation. The purpose of the second qualitative experiment was to discuss navigation with users and gather their opinion on enhancement of navigation in the domain by exploiting their navigation history in the cloud. Lastly, we conducted a quantitative experiment with all users in Annota (active at the time of experiment) during one-week period of time.

A. Experiment 1: Comparison of user's queries

We divided 4 users into two separate groups. The first group (Group 1) had the interface without terms from history and the other group (Group 2) was using interface with history and its visualization. Since users were interacting with the system for the first time, we chose exploratory search as a use case for exploiting navigation history of all users in a day-long period of time. Feedback from users was acquired by a questionnaire and mutual discussion at the end of our experiment.

We asked users to navigate in order to get familiar with two topics in the dataset. The topics were chosen according to the nature of the dataset:

1) Information retrieval and experiments concerning information retrieval
2) Navigation and data used in navigation

After completing the navigation tasks, we analysed gathered data consisting of user's queries. Tab. I summarizes queries constructed by groups during navigation.

According to the logs gathered while navigating, we conclude that users in Group 2 were navigating more extensively and their information need was adapting as their knowledge of information space enhanced. The extensive usage by Group 2 was caused by the fact that each member of the group could see the query that the others constructed similar to theirs.

In the questionnaire we asked the users from the Group 2 to answer questions regarding our cloud navigation and its interface. We asked them the following questions:

1) **Describe your motivation for choosing a colored word from cloud.**
   The participants agreed upon the notion that the words highlighted with color were more dominant in

[2]http://www.elasticsearch.org/
[3]http://www.alchemyapi.com/

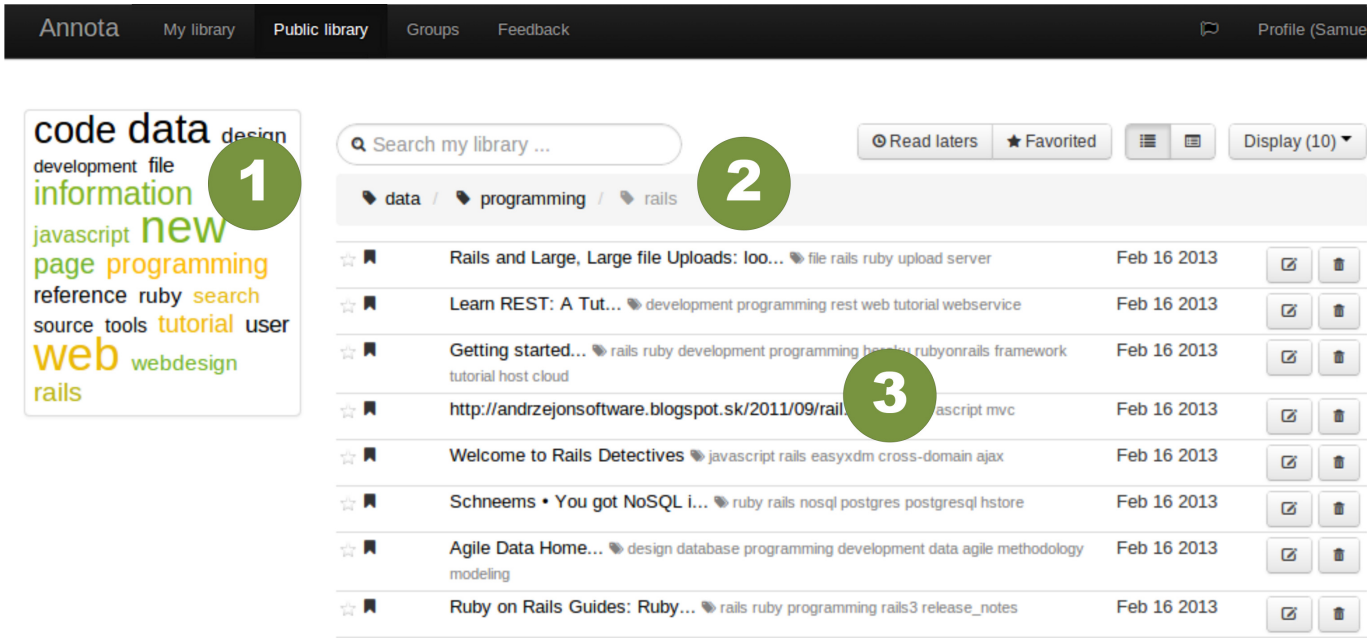

Fig. 1. Term cloud navigation in Annota. The users can navigate using term cloud (1), in which the terms from history are differentiated by color (from oldest to newest as an interval from orange to green). The query is represented as a sequence of terms selected from the cloud (2). The results are ranked considering the position of words in the query (3).

TABLE I. USERS' QUERIES AND THEIR USAGE DURING NAVIGATION.

Topic 1			
Group 1		Group 2	
Term	Count	Term	Count
information	5	information retrieval	12
information retrieval	3	paper	6
retrieval	2	experimentation	5
search process	2	experiment	3
		information search	2
		retrieval	1
		theory	1
Topic 2			
Group 1		Group 2	
Term	Count	Term	Count
navigation	5	data	13
data	5	navigation	14
data mining	1	design	3
		semantic web	2
		data mining	1
		web search	1

the cloud and therefore they were more focused on finding relevant words among the highlighted ones.

2) **Rank the relevancy of the resultant documents.** The mutual agreement was that the documents at the top of the result list were more relevant than the documents at the bottom and they noticed that the result list was adapting to the position of the terms in the query.
3) **How did highlighted words effect your navigation?** The mutual agreement was that participants chose highlighted words, because they were both visually distinguished and relevant to their query. Users agreed that the color was visually more dominant than font size and they used smaller highlighted words in favour of bigger, but colorless ones.

According to the results of our experiment and answers from the questionnaire, our method assisted users during exploratory navigation and helped them to refine their queries by exploiting queries of other users. The participants positively reacted on query refinement using highlighted words, but in the beginning they had a difficulty understanding what the colour meant. The results of the experiment might be more accurate if the navigation history was more extensive in the selected time period, therefore the cloud would provide more navigation paths to follow.

B. *Experiment 2: Qualitative evaluation with users*

The goal of the second experiment was to observe how the users interact with the proposed method of navigation and whether it is useful for them. Again, we conducted the experiment with 4 participants. As a dataset we used 4313 documents consisting of website bookmarks and articles from various digital libraries. Fulltext search was included as a part of navigation interface. Participants were supposed to navigate to articles concerning the following two topics (represented as a list of terms):

1) Navigation, tags, experiments and web
2) Programming, algorithms and web

During navigation we tracked the participants' queries and the content of the cloud was extended with terms from the queries, but terms were not highlighted by different colors. At the end of the experiment we highlighted terms in the cloud by color to emphasize their relevance in period of time and participants expressed their opinion what the color meant. The mutual agreement was that the color is a from of recommendation of relevant terms to their current query and they paid more attention to highlighted words with smaller font size than bigger ones without specific color shade, since they believed that highlighted terms were more relevant based on system recommendation. The users did not intuitively

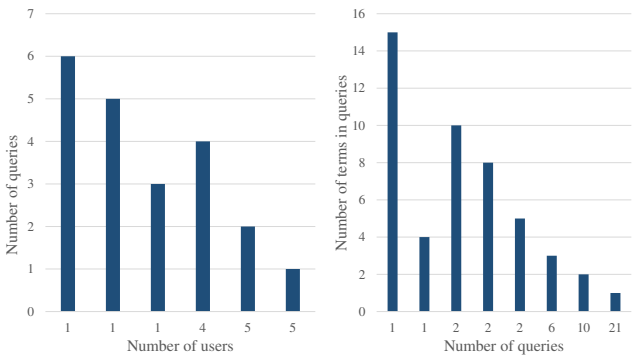

Fig. 2. Distribution of navigation data by the number of users' queries (left) and the number of terms in queries (right).

recognize that different shades of color represent last usage of words in history.

We also noticed that the majority of the users navigated by similar pattern. Firstly, they searched for term using fulltext search and then they used cloud to navigate among the resultant documents. The motivation for this pattern was that after specifying exact query, they filtered documents that are relevant to their information need and the content of the cloud consisted only of relevant terms to their query.

Based on the findings from this experiment, we can conclude that users are receptive to different term colors more than to their size. In addition, if they are faced with a navigation task, they tend to filter the information space using the fulltext search and then navigate among the results using term cloud to refine the query and explore the filtered subspace in more detail.

C. Experiment 3: Quantitative evaluation with Annota users

We conducted the last experiment with 17 users of Annota during one week. The content of cloud was extended with words from history without highlighting. The utilization of colors to represent history was not part of the experiment, since the goal of the experiment was to quantitatively evaluate proportion of term selection from history with respect to all the terms in the cloud.

We gathered navigation data consisting of number of terms in cloud, number of terms selected from documents to cloud and number of terms selected from history to cloud. In addition, we tracked which documents users chose during navigation (e.g. user clicked on a document link) and their position in the result list.

The navigation data showed that 17 users who navigated using cloud in Annota during our experiment constructed 45 queries consisting overall from 124 terms (65 unique). Fig. 2 shows distribution of navigation data by number of queries and number of terms in queries.

We evaluated the proportion of terms from history selected from the cloud by the users and proportion of terms from history present in the cloud. We calculate the latter as the ratio of the terms in the cloud added from navigation history to all the cloud terms in every navigation step:

TABLE II. SELECTION AND PROBABILITY OF SELECTION OF TERMS ADDED FROM HISTORY INTO THE CLOUD.

N	History	Random	Difference
1	0.40	0.07	0.33
2	0.16	0.05	0.11
3	0.26	0.09	0.17
4	0.14	0.13	0.02
5	0.00	0.23	-0.23
6	0.30	0.27	0.03
7	0.18	0.20	-0.02
8	0.35	0.19	0.16
9	0.44	0.27	0.18
10	0.55	0.26	0.29

$$W_h = \frac{\sum_{\forall term\ selection} |history\ terms\ in\ the\ cloud|}{\sum_{\forall term\ selection} |all\ terms\ in\ the\ cloud|} \quad (5)$$

Besides the proportion of words in cloud from history, W_h defines a probability with which the users can select any term from the cloud into their queries. We also calculate the actual frequency of selecting the terms from history by the users as a fraction of the number of terms added from history that were selected by the users to their queries and the number of all the selected terms from queries.

$$Q_h = \frac{|selected\ terms\ added\ from\ history|}{|selected\ terms|} \quad (6)$$

The users selected 124 terms (65 unique) in their queries. According to (5) and (6) and the data we gathered from the users' navigation sessions, $W_h \approx 0.14$ and $Q_h \approx 0.30$. It means that the clouds consisted from 14% of terms from history on average during each navigation step. On the other hand, about 30% of users' queries consisted of terms from history. Based on the presented results, we can conclude that the users selected terms from history twice as often as they would if they selected terms in the cloud randomly. Therefore, it suggests that terms added into the cloud from history were deemed relevant by the users.

In order to further support our findings, we examined the relationship between the frequency of selection of cloud terms added from the navigation history and the probability of selecting the history terms if the users would have selected the terms during their navigation sessions randomly with respect to the overall number of terms added from history present in the cloud. Tab. II shows the number N of the terms added from history present in the cloud in each navigation step. *History* is represented by Q_h from (6) and *Random* by W_h from (5). *History* means that user has selected term added from history into the cloud to her query on purpose and *Random* represents probability of terms being selected by the user to the query.

As we can see in Fig. 3 in the majority of the cases the actual selection frequency of the history terms was higher than the probability of their selection by chance. For example, with only one history term in the cloud, the probability of its selection was about 7% and yet it was selected by the users 40% of the cases. On the other hand, with 5 history words in the cloud, we did not observe any selection of these terms by the users. It could have been caused by the relatively short duration of our experiment and the size of the collected

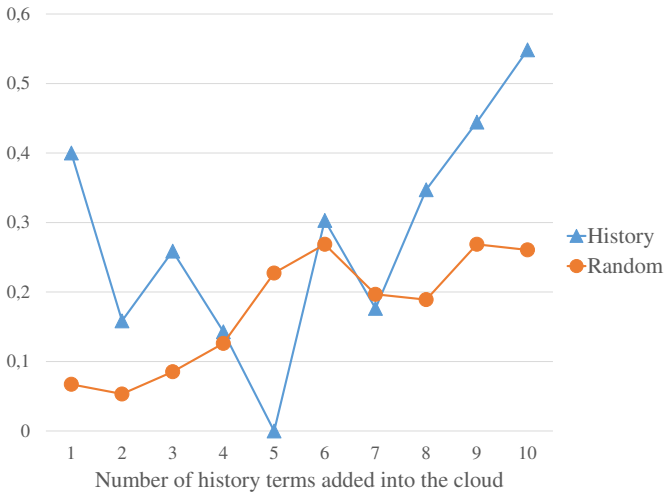

Fig. 3. Selection and probability of selection of term added from history into the cloud by history term count distribution in the cloud.

data. Therefore, we decided to use a statistical test in order to determine whether the observed differences are relevant.

We used Wilcoxon signed-rank test for this purpose, since the data population was not normally distributed. Values of *History* and *Random* from Tab. II were used as pairs. Based on the results we can reject the null hypothesis $H0$ (that the difference of medians between the paired sets is zero) at the significance level of $p-value = 0.05$. Hence, we can conclude that that the observed differences are relevant and the users selected terms from history more frequently even when they were not distinguished by the color. This strongly suggests that these terms are valuable for the users when navigating in the information space.

V. Conclusion

Our contribution towards cloud navigation is in utilization of history in a period of time to provide extended content of term cloud with color-based visualization of history. We utilize navigation history of all users to enhance navigation of novice users. The history records provide an overview of the users' interests in the particular period of time by frequency of term occurrences. We extend the content of term cloud with terms that are similar to current user's query by finding similar queries from history. The different shades of colors distinguish last usage of term in history relative to the current query.

We conducted a series of experiments to evaluate our method when users explore a domain and their information need is more general. Based on the results of our evaluation we conclude that visualisation of history of all the users is beneficial for the users' query refinement and their first steps in exploring new domain. The quantitative experiment in Annota showed that users refine their queries with terms added from history, because the terms are relevant to their query.

Although we have focused on the scenario of exploratory navigation of novice researchers in the digital libraries domain, the proposed method can be modified in order to support revisitation and refinding scenario as well. In that case we would consider only history of one user instead of history of all users. We plan to examine the effect of utilizing only the subset of users and their navigation history, e.g. consider only the users' with similar interests or on the contrary, the users' which can provide the given user with new topics to explore in order to maximize diversity. This could also help us tackle the filter bubble problem, which can occur when the users select only recommended terms from the history.

In addition, we plan to implement more complex similarity function to match similar queries and documents when constructing cloud content, since current similarity function consider query and document as similar if it consists of at least one term in user's query. Lastly, extended quantitative experiment is planned in the future with the aim to verify other aspects of the proposed method, such as the relevancy of the retrieved documents and their ranking.

Acknowledgment

This work was partially supported by the grant No. VG1/0971/11 and the grant No. APVV-0208-10. The authors also wish to thank all the members of the PeWe group (pewe.fiit.stuba.sk) for their invaluable contribution to discussions and support in experiments.

References

[1] J. Gwizdka, L. Studies, N. Brunswick, and P. Bakelaar, "Tag Trails: Navigation with Context and History," *Design (2009)*, vol. 69, no. 2, pp. 4579 – 4580, 2009.

[2] O. Kaser and D. Lemire, "Tag-cloud drawing: Algorithms for cloud visualization," *CoRR*, vol. abs/cs/0703109, 2007.

[3] G. Koutrika, Z. Zadeh, and H. Garcia-Molina, "Data clouds: summarizing keyword search results over structured data," in *Proc. of the 12th Intl. Conf. on Extending Database Technology Advances in Database Technology*, 2009, pp. 391–402.

[4] C. S. Mesnage and M. J. Carman, "Tag navigation," in *Proc. of the 2nd Intl. Workshop on Social Software Engineering and Applications - SoSEA '09*. New York, New York, USA: ACM Press, 2009, p. 29.

[5] P. Návrat, "Cognitive traveling in digital space: from keyword search through exploratory information seeking," *Central European Journal of Computer Science*, vol. 2, no. 3, pp. 170–182, 2012.

[6] K. Rástočný, M. Tvarožek, and M. Bieliková, "Supporting Search Result Browsing and Exploration via Cluster-Based Views and Zoom-Based Navigation," in *Proc. of the 2011 IEEE/WIC/ACM Intl. Conf. on Web Intelligence and Intelligent Agent Technology - Volume 03*, ser. WI-IAT '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 297–300.

[7] G. Solskinnsbakk and J. A. Gulla, "Contextual search navigation using semantic tag signatures," in *Proc. of the 11th Intl. Conf. on Knowledge Management and Knowledge Technologies - i-KNOW '11*. New York, New York, USA: ACM Press, 2011.

[8] J. Suchal and P. Návrat, "Full Text Search Engine as Scalable k-Nearest Neighbor Recommendation System," in *Proc. of IFIP Advances in ICT, Held as Part of WCC 2010*, vol. 331, 2010, pp. 165–173.

[9] J. Ševcech, "Related Documents Search Using User Created Annotations," *Information Sciences and Technologies Bulletin of the ACM Slovakia*, vol. 5, no. 2, pp. 44–47, 2013.

[10] J. Ševcech, M. Bieliková, R. Burger, and M. Barla, "Researcher activity tracking in digital library of scientific resources enriched by annotations. (in Slovak)," in *Proc. of 7th Workshop on Intelligent and Knowledge Oriented Technologies - WIKT 2012*, 2012, pp. 197–200.

[11] M. Šimko, "Automated Acquisition of Domain Model for Adaptive Collaborative Web-Based Learning," *Information Sciences and Technologies Bulletin of the ACM Slovakia*, vol. 4, no. 2, pp. 1–9, 2012.