Linking Related Documents: Combining Tag Clouds and Search Queries

Christoph Trattner and Denis Helic

Graz Technical University of Technology
Inffeldgasse 21a/16c
A-8010 Graz
ctrattner@iicm.edu, dhelic@tugraz.at

Abstract. Nowadays, Web encyclopedias suffer from a high bounce rate. Typically, users come to an encyclopaedia from a search engine and upon reading the first page on the site they leave it immediately thereafter. To tackle this problem in systems such as Web shops additional browsing tools for easy finding of related content are provided. In this paper we present a tool that links related content in an encyclopaedia in a usable and visually appealing manner. The tool combines two promising approaches – tag clouds and historic search queries – into a new single one. Hence, each document in the system is enriched with a tag cloud containing collections of related concepts populated from historic search queries. A preliminary implementation of the tool is already provided within a Web encyclopaedia called Austria-Forum.

Keywords: query tags, tags, tag clouds, linking.

1 Introduction

Content in Web encyclopedias such as Wikipedia is mainly accessed through search engines. Typically, users with an interest in a certain encyclopedic topic submit a Google search, click on a Wikipedia document from the result list and upon reading the document they either go back to Google to refine their search, or close their browsers if they have already found the information they needed. Such a user behaviour on encyclopedia sites is traceable through a typical high bounce rate (see Alexa[1] for instance) of such sites. Essentially, users do not browse or search in Wikipedia to find further relevant content - they are rather using Google for that purpose.

In our opinion, Web encyclopedias lack simple and usable tools that involve users in explorative browsing or searching for related documents. In other Web systems, most notably Web shops, different approaches have been applied to tackle this situation. For example, one popular approach involves offering related information through collaborative filtering techniques as on Amazon. Google or Yahoo! apply a similar approach to offer related content by taking the users' search query history into account via sponsored links [4].

[1] http://www.alexa.com/siteinfo/wikipedia.org

B. Benatallah et al. (Eds.): ICWE 2010, LNCS 6189, pp. 486–489, 2010.

Recently, social bookmarking systems emerged as an interesting alternative to search engines for finding relevant content [3,7]. These systems apply the concept of social navigation [5] i.e. users browse by means of so-called tag clouds, which are collections of keywords assigned to different online resources by different users [2] driven by different motivations [8].

In this paper we introduce a novel approach of offering related content to users of Web encyclopedias. The approach is based on simple idea of integrating a tagging system into a Web encyclopedia and populating that system not only with user-generated tags but also with automatically collected Google query tags. In this way we combine two promising approaches successfully applied elsewhere into a single new one – the access to related content is granted not only through social trails left in the system by other users but also through search history of general user population. To test this idea a prototype tool has been implemented within a Web encyclopedia called Austria-Forum[2].

The paper is structured as follows. Section 2 presents the basic idea of this new approach and provides an analysis of its potentials. Section 3 shortly discusses the implementation of the idea within Austria-Forum. Finally, Section 4 concludes the paper and provides an outlook for the future work in this area.

2 Approach

The basic idea of this new approach is to combine provision of related documents as offered by social bookmarking sites and by e.g. Google search query history. On the one hand, tag clouds represent a usable and interesting alternative navigation tool in modern Web-based systems. Moreover, they are very close to the idea of explorative browsing [6], i.e. they capture nicely the intent of users coming to a system from a search engine - users have searched in e.g. Google and now they click on a concept in a tag cloud that reflects their original search intent. On the other hand, Google search query history, i.e. queries that are "referrers" to found documents are an invaluable source of information for refining user search in the system. It is our belief that an integration of such historical queries into a tag cloud user interface provides a promising opportunity to lead users to related documents.

To make this idea work the tag clouds need to be calculated in a context or resource-specific way, i.e. each resource in the system is associated with a special tag cloud. This resource-specific tag cloud captures the most important concepts and topics related to the current document and hence provides a useful navigational tool for exploration of related resources in the system. In addition to the user-generated tags the related concepts and topics are obtained from historic Google search queries leading to the resource in question.

To investigate the feasibility of this approach before implementing it, we conducted an analysis of tagging data automatically obtained from Google queries for Austria-Forum (AF). Thus, Google query tags have been collected over a period of four months and analyzed using the following metrics: number of tags

[2] `http://www.austria-lexikon.at`

Table 1. Growth of tagging set over time with user-generated and Google query tags

(a) User Tags

day	$\#t$	$\#t_{new}$	$\#r$	$\#r_{new}$
-200	3,202	3,202	4,884	4,884
-160	7,829	4,627	7,450	2,566
-120	8,980	1,151	9,109	1,659
-80	10,009	1,029	11,523	2,414
-40	10,628	619	12,421	898
now	11,097	469	12,871	450

(b) Google Query Tags

day	$\#t$	$\#t_{new}$	$\#r$	$\#r_{new}$
-60	3,906	3,906	1,698	1,698
-50	7,020	3,114	3,160	1,462
-40	10,018	2,998	4,710	1,550
-30	12,772	2,754	6,245	1,535
-20	15,615	2,843	8,055	1,810
-10	17,743	2,128	9,368	1,313
now	19,867	2,124	10,659	1,291

$\#t$, number of new tags $\#t_{new}$, number of resources $\#r$, and number of new resources $\#r_{new}$. The analysis observed the changes in these metrics over time.

Table 1 shows the potential of the Google query term approach. Over 10,659 AF resources were tagged during a period of 60 days by Google query tags. Compared to the user-generated tags in AF, which show an average increase of 399.35 tagged resources per 10 days for the last 200 days (see Table 1(a)), an average of around 1,500 new tags per 10 days for the last 60 days has been achieved with the query tags (see Table 1(b)). Thus, automatic tagging approach annotated four times more resources than the human approach within AF.

As the last step two tagging datasets have been combined. The combined dataset annotates 20,688 resources, which is an increase of nearly 100% in the number of annotated resource as compared to user-generated tags. Additionally, the combined dataset contains 27,824 unique tags (an increase of 150%). Similar results have been obtained by [1] for the `stanford.edu` domain.

3 Implementation

The first prototypical implementation of the tool consists of four modules.

Tag Collection Module: The module consists of two sub-modules: a client- and a server sub-module. The client collects HTTP-Referrer information of the users that comes from the Google search engine to a particular resource within Austria-Forum. The client sub-module is implemented via JavaScript AJAX. The server is a Web service that processes HTTP-Referrer headers sent over by the client sub-module. Firstly, the service identifies single query terms and denotes them as potential *tags*. Secondly, to filter out the noisy tags a stop word and a character filter is applied.

Tag Storage Module: This module stores the tags obtained by the collection module into a database. Currently, the tag database is hosted on a MySql server as a normalized tag database.

Tag (Cloud) Generation Module: To provide the access to related documents a resource-specific tag cloud is calculated by this module. This tag cloud is of the form $TC_r = (t_1, ..., t_n, r_1, ..., r_m)$, where $r_1, ..., r_m$ are the resources that have any of $t_1, ..., t_n$ tags in common. The calculated tag clouds

are cached on the server-side to improve the performance of the system. For retrieving the tags and the corresponding resources this module provides a simple interface that consists of two functions: $GetTags(URL)$ (generates a XML representation of a tag cloud), and $GetLinks(URL, tag)$ (generates a XML representation of the resource list for a particular tag).

Tag Cloud Presentation Module: This modul is a client-side AJAX module implemented in JavaScript. It manipulates the browser DOM objects to render a tag cloud in a visually appealing fashion.

4 Conclusions

In this paper we presented an approach for exploring related resources in Web encyclopedias. The tool aims at offering additional navigational paths to related resources for users of such systems in general, and for users who come to these systems from a search engine such as Google. The future work will include development of a theoretical framework to compare this approach to other approaches aiming at a provision of related content in web-based information systems. In addition to theoretical investigations, a usability study to assess the acceptance and usefulness of the tool will be carried out.

References

1. Antonellis, I., Garcia-Molina, H., Karim, J.: Tagging with queries: How and why. In: ACM WSDM (2009)
2. Heymann, P., Paepcke, A., Garcia-Molina, H.: Tagging human knowledge. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, New York, NY, USA, pp. 51–61 (2010)
3. Mesnage, C.S., Carman, M.J.: Tag navigation. In: SoSEA '09: Proceedings of the 2nd International Workshop on Social Software Engineering and Applications, New York, NY, USA , pp. 29–32(2009)
4. Mehta, A., Saberi, A., Vazirani, U., Vazirani, V.: AdWords and generalized online matching. J. ACM 54(5), 22 (2007)
5. Millen, D., Feinberg, J.: Using social tagging to improve social navigation. In: Workshop on the Social Navigation and Community Based Adaptation Technologies, Dublin, Ireland (2006)
6. Sinclair, J., Cardew-Hall, M.: The folksonomy tag cloud: when is it useful? Journal of Information Science, 34–15 (2008)
7. Strohmaier, M.: Purpose Tagging - Capturing User Intent to Assist Goal-Oriented Social Search. In: SSM'08 Workshop on Search in Social Media, in conjunction with CIKM'08, Napa Valley, USA (2008)
8. Strohmaier, M., Koerner, C., Kern, R.: Why do Users Tag? Detecting Users' Motivation for Tagging in Social Tagging Systems. In: 4th International AAAI Conference on Weblogs and Social Media (ICWSM 2010), Washington, DC, USA, May 23-26 (2010)