

Classification of User Interest Patterns Using a Virtual Folksonomy

Ricardo Kawase
L3S Research Center
Hannover, Germany
kawase@L3s.de

Eelco Herder
L3S Research Center
Hannover, Germany
herder@L3s.de

ABSTRACT

User interest in topics and resources is known to be recurrent and to follow specific patterns, depending on the type of topic or resource. Traditional methods for predicting reoccurring patterns are based on ranking and associative models. In this paper, we identify several ‘canonical’ patterns by clustering keywords related to visited resources, making use of a large repository of Web usage data. The keywords are derived from a ‘virtual’ folksonomy of tags assigned to these resources, using a collaborative bookmarking system.

Categories and Subject Descriptors

H.5.4 [Hypertext/Hypermedia]: User Issues

General Terms

Human Factors

Keywords

Revisitation, Navigation support, Recommendation

1. INTRODUCTION

Many of the places that we visit on the Web are places that we visited before. The majority of revisits is covered by a small number of popular places - such as the user’s favorite search engine, online retailers, social networking and news sites - and places visited in the very recent past [6]. The same power law distribution can be observed in the overall popularity of Web sites, friend connections in social networking sites and tag collections [3]. Search engines and recommender systems have exploited these regularities for several decades [4].

In the past few years, researchers have shown an increasing interest in the long tail that follows the top-k most popular, most frequent or most recent sites, queries, tasks and connections [8]. The reason: even though the few items in the top-k account for a proportionally large proportion of user behavior, most time and attention is accounted for by the many remaining items in the long tail.

Current browser history support - most prominently the back button, bookmarks and url autocompletion - has been shown not to

sufficiently support revisits to the long tail [11, 2]. These long-term revisits represent people’s long-term (niche) interests, infrequent information needs, entertainment and hobby activities. Often, users need to re-search and retrace a site’s name or address, or forget about its very existence. For this reason, the analysis and prediction of online browsing behavior has received much attention from research and industry [2, 15, 10, 7, 12].

In a related branch of research, the *folksonomy* of tags given by users to various resources [9] is successfully exploited to build tag-based profiles of both users and resources. These profiles are used for personalization [5], recommendation and improvement of search [3].

The starting point of the research discussed in this paper is the observation that recurrent activities on the Web represent recurrent user interests, tasks and goals; many of the revisited resources have been annotated with tags by various users, and these tags represent the ‘wisdom of the crowd’ on what these resources are used for. This public folksonomy is assumed to be more representative than the user’s individual tags - which are often subjective [16] and low in number - or the keywords in the page title.

In this paper, we aim to identify and explain ‘canonical’ patterns of reoccurring activities based on tag occurrences in the users’ on-line lives. We do this by relating client-side Web usage logs with the tags that describe the resources in these logs. We developed a classification of the most common patterns of user interest by clustering keywords by their appearance on the users’ timelines. These patterns vary from one-time interest to repetitive peaks and constant interest. An analysis of the top keywords related to these patterns shows that these patterns differ from one another in terms of user interests, tasks and goals. To prove that it is feasible to use the classification for automatically recognizing these patterns, we implemented and evaluated a simple rule-based heuristic classifier.

The remainder of this paper is organized as follows. After a short discussion of related work, we formally define the *virtual folksonomy* and the data sets that were used to build this folksonomy. In Section 4, we present the results of the clustering and classification methods applied to the data. We conclude with a discussion of our findings in section 5.

2. RELATED WORK

There is a large body of studies on recurrent behavior on the Web, the first one being carried out by Tauscher and Greenberg [14] late 1995. Obendorf et al. [11] distinguished *short-term revisits* (backtrack or undo) from *medium-term* (re-utilize or observe) and *long-term revisits* (rediscover, reuse). Further, different categories of sites invite different revisit behavior: search engines typically have one page that users frequently return to, institutional sites also comprise a long tail of pages visited several times.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

JCDL’11, June 13–17, 2011, Ottawa, Ontario, Canada.

Copyright 2011 ACM 978-1-4503-0744-4/11/06 ...\$10.00.

Adar et al. [2] found that short-term revisits involve hub-and-spoke navigation, visiting shopping or reference sites or pages on which information was monitored. Medium-term revisits involve popular home pages, Web mail, forums, educational pages and the browser homepages. Long-term revisits involve the use of search engines for revisitation, as well as weekend activities such as going to the cinema. A subsequent study was carried out [15] and confirmed that short-term revisits mainly involves continuing working on a task or routine behavior, whereas longer-term revisits mainly involves re-evaluation and reuse.

Various types of annotation mechanisms exist for relating keywords to Web resources. Most HTML documents have a title and metatags that are made by the creator; Web directories may provide catalogue information; keywords may be automatically extracted from the text; or they may be provided by the user [16]. The word *tag* typically refers to user-generated keywords. As tags are not primarily meant for labeling resources, not all of these tags are (topic-related) descriptions of the resource: tags may also represent the original taggers’ tasks and personal opinions. Carmagnola et al.[5] investigated the nature of tags in a touristic information site and found that the overall majority of tags were topic-related; one-third of the tags contained additional information, such as category, resource context, and synonyms. Bischoff et al.[3] investigated to what extent tags can be used for search and found that about 50% of tags assigned to resources in Delicious¹ are topic-related keywords: non-subjective annotations that relevant for all users. Further, there is a significant overlap of the keywords used in tags and those used in queries.

3. GENERATING A VIRTUAL FOLKSONOMY

A traditional *folksonomy* is a quadruple $\mathbb{F} := (U, T, R, Y)$, where U, T, R are finite sets of instances of *users*, *tags*, and *resources*, respectively. Y defines a relation, the *tag assignment*, between these sets, that is, $Y \subseteq U \times T \times R$, possibly enriched with a timestamp that indicates *when* it was performed [9].

We created a *virtual folksonomy* by enriching the a client-side Web usage log - which contains Web pages (R) that are visited by users (U) - with tags (T), making use of the social bookmarking system Delicious. We call the folksonomy ‘virtual’ because of the indirect manner in which tags are associated with the individuals’ Web histories.

The Web History Repository² is a public repository of anonymized Web usage data that researchers can use to gain new insights in online browsing behavior. The data includes the list of visited pages, including timestamp and browser session. For each visited page, the (encrypted) url and host, the total number of visits, the frequency and the last visit is listed in a separate table. The repository contains data of 201 users, with a total of 1,324,041 visits to 857,271 unique URLs.

We crawled the online bookmarking system Delicious to retrieve the user-provided tags associated with each URL, thus enriching the web usage log into a virtual folksonomy. In total, we found 10,696 unique URLs that have been tagged with 331,699 tags, summing up to a total of 64,179 unique tags. As expected, Delicious contained tags for only a subset of the pages in the Web usage logs. Still, these pages accounted for 7% of the total number of page visits and thus sufficiently covers the long tails in the user’s logs.

As analyzed by [1], the combination of user-specific usage data with popular tags is an effective mechanism for improving the per-

¹<http://www.delicious.com/>

²<http://webhistoryproject.blogspot.com/>

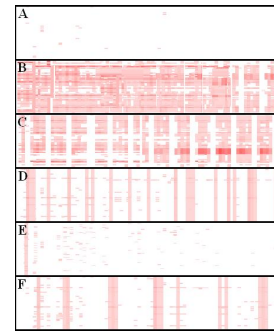


Figure 1: The different clusters plotted by Cluto. Each row represents a cluster; darker colors represent a higher number of occurrences

formance of tag recommendations, in particular during the cold-start period, when little or nothing is known about the user. Further, apart from enhancing incomplete profiles, it is a method for diversifying the profile semantics by combining a user’s specific behavior with the wisdom of the crowd.

4. TAG-BASED INTEREST PATTERNS

In this section, we focus on the identification of tag-based user interest patterns. As a first step, we clustered the tag revisitation curves based on the similarities with respect to time; we use the most common keywords associated with each cluster to explain its meaning. Second, based on the general shapes of the clusters, we developed a rule-based classifier that maps each keyword to the groups derived from the cluster. At the end of the section, we discuss the characteristics of the interest patterns found.

4.1 Clustering Interests

In order to identify ‘canonical’ patterns of recurrent user interests, we followed the clustering and classification approach introduced by Adar et al. [2], who used it for evaluating revisitation behavior for different URLs. The clustering output is a normalized revisitation curve that identifies different types of revisitation.


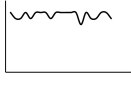
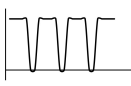

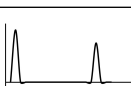

In our experiment, the URLs are translated to the tags associated with the URL in the virtual folksonomy. Due to the overlap of keywords between URLs, the curves represent generic user interests rather than reuse of specific pages or sites. As we are interested in longer-term patterns, we group the keywords of interest in buckets of each one day. To align differences in starting point and time span covered by the logs between users, we employed several normalization strategies as used by Yang and Leskovec [17].

All data was aligned by shifting all first keyword appearances to a ‘point zero’, all further appearances of this keyword were shifted to the corresponding distances from this point zero. To observe weekly routines, we preserved the weekday information during the shifting process: for example, a curve of interest that started on a Tuesday in the second month of a user’s history is shifted to start on Tuesday in ‘week zero’. We did not normalize the time span of the keyword life times, as techniques such as Dynamic Time Warping would introduce artificial patterns due to the stretching.

With the aligned data, we used repeated-bisection clustering with a cosine similarity metric [13]³. Varying the similarity metrics and the number of clusters, we found six well-defined clustered behaviors, as depicted in Figure 1. We manually analyzed these clusters, named them based on general trends and summarized these trends with descriptions and example keywords, as depicted in Table 1. It

³<http://glaros.dtc.umn.edu/gkhome/views/cluto/>

Table 1: Summarization of Web users' interest.

Group	Shape	Description	Examples
C1 - One-time interest.		This group represents an interest that happens a single time during the user's history. This includes spam, involuntary access and typos.	condos, gotcha, ebuddy, job, googlemaps
C2 - Constant interest.		This group shows high a constant interest of the user in a topic or service: Search engines, news papers, webmail.	yahoo, google, search, email, magazine, news, sports, portal,web, daily, community
C3 - Constant interest with repetitive drops.		This group represents an constant interest with repetitive drops, mostly caused by weekend breaks. Similar to constant interest in a working environment.	yahoo, google, search, email, magazine, news, sports, portal,web, daily, community
C4 - Repetitive peaks.		This group represents regular, repetitive peaks of interest, mostly caused by exclusive weekend accesses and weekly routines. Websites of games, sports, TV shows, regular meetings.	institute, project, download, documents, sport, channel, games, fun, cosmology, soccer, streaming.
C5 - Sporadic standalone peaks.		This group contains interests that return on an irregular basis and do not last longer than a day. This includes finance, specialized reference sites, restaurant finding.	movies, restaurant, maps, banking, imdb, xml, java.
C6 - Sporadic connected peaks.		This group shows interests that return on an irregular basis and that typically last longer than a day, such as online shopping, travel planning and research activities.	party, ebay, airline, trip, mathematics, ask, java, .net, humor, learning, wikipedia, wiki, research

is worth noting that the descriptions are derived from our qualitative analysis of the most representative tag-examples found in each cluster.

4.2 Classifying Interests

Following the clustering process, we implemented a rule-based heuristic classifier that assigns a keyword to one of the groups that correspond to the identified clusters, based on the keyword's occurrence pattern. With the classifier, we aim to verify the usefulness of the classification derived from the clusters, in terms of discriminative power. Further, the distribution of keywords in the groups is expected to provide insight in temporal dynamics of user interests.

During the classification process, we recognized a missing pattern that was not clearly identified during the data clustering, due to few occurrences and similarities with other clusters. The 'missing' canonical curve represents interests that happened during a continuous period of time in the users' history but that never pops up again (C7), as depicted on the top right of Figure 2.

Once the seven classifications were defined, we implemented a mutually exclusive classifier based on a set of rules that identify the canonical curves. In other words, each user's interest belongs to one, and only one group. The classifier incrementally iterates over the whole array of occurrences of a tag and, for each iteration, it assigns the possible group to the tag. This implementation allows us to incrementally verify the classification changes of each tag over time and also supports streaming data (as is the case of browsing history in real use). The rules can be summarized as follows:

- C1 - a keyword is used on one single day
- C2 - a keyword is used during a longer consecutive period, containing only a few days on which the keyword is not used
- C3 - a keyword is used during the whole logging period, containing several days on which the keyword is not used; gaps between these days are evenly distributed
- C4 - a keyword is used on a regular basis, low deviation of gaps between appearance

- C5 - a keyword is used on a regular basis, high deviation of gap length between appearances, peaks last only a day
- C6 - similar to C5, but the peaks of keyword appearance last longer than a day
- C7 - a keyword is used in one single period of more than one consecutive days

We evaluated the classifier on the virtual folksonomy, as described in Section 3. To avoid bias introduced by popular URLs with many tags (such as the Google portal page), we only considered the top-10 tags per URL. Second, since we are interested in modeling user interest based on a long term history we ignored all users that had less than 28 days of history. The resulting dataset consisted of 71 users and 8095 tags representing these users' interests.

4.3 Classification Results

The distribution of the users' interests is exposed in Figure 2. The sizes of the bars represent the number of keywords assigned to a group; the line indicates the number of page accesses related to the keywords in this group.

A first observation is that the majority of keywords (around 55%) is used only once during a single day (C1). However, the page accesses related to these keywords covers less than 5% of the users' access logs. By contrast, less than 1% of the users' keywords is used on a daily or very constant basis (C2 and C3), but these groups cover with 28% a large portion of the users' accesses. In other words, C2 and C3 represent the head of the power law distribution of keyword usage, C1 the very end of the tail.

Groups C4, C5 and C6 represent the middle part of distribution, covering 40% of the users' interests and 65% of the page accesses. These groups are associated with the users' repetitive and sporadic interests. Group C4 confirms the existence and importance of routine weekly interests. Still, most page visits concern group C6 - the sporadic peaks of interest that last more than one day. This implies that irregularly returning tasks - such as online shopping,

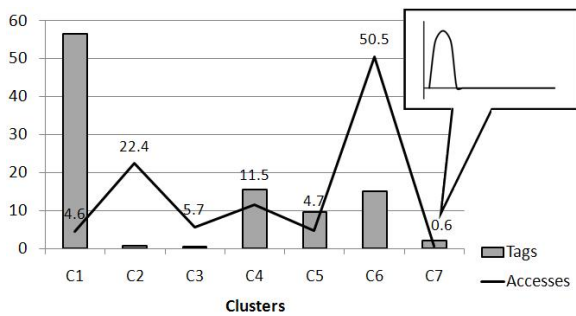


Figure 2: Y-Axis shows the distribution of the tags in clusters (bars) and the number of page visits covered by the group (line).

travel booking and background research - result in far more revisits than daily or constant interests (C2, C3) and that these returning tasks typically last longer than one day. Finally, the ‘unidentified’ cluster C7 (one-time interests that cover several days) contains very few keywords and page accesses. We assume that this indicates that the lifetime one-time interests is typically short (C1); if the interest remains longer than one day (C7), most likely it will happen again (C4, C5 or C6).

5. DISCUSSION AND CONCLUSIONS

In this paper, we analyzed patterns of returning user interests, making use of a *virtual folksonomy*, composed of client-side web logs enriched with social bookmarking tags.

The results indicate that the greater part of user interests involves tasks that turn up on a more or less regular basis and typically involve longer-lasting activities as travel planning and (goal-directed) shopping activities. If an interest remains longer than one days, it is likely to return at a later stage.

The results are based on a folksonomy, representing about 200 users, 1 million page visits and 65,000 tags. These large numbers and the similarity metrics of the clusters suggest that the results are representative for the ‘average’ web user. However, the interpretation of the clusters and the choices made during the creation of the classifiers are likely to have caused some bias in the quantities of the results. We are confident that this bias did not have an impact on the trends described in this paper.

The dominance of the middle part of the power law distribution of keywords is yet another plea to reduce the dominance of the most frequent items and focus on the (start of the) tail instead. In the context of Web browsing, this middle part is mainly formed by interests that return on a more or less regular basis. These patterns of temporal variation can be exploited to better relate keywords, tags or other items in a user profile. Many applications can be thought of in the context of personalization and recommender systems, such as repetitive-interest based collaborative filtering or product recommendations.

A further application area, which we aim to further explore, is the prediction of page revisits - which can be used for suggestions in browser toolbars or portal sites, and personalization of search results. Some straightforward applications include reminders for regularly repeating activities, or contextual revisitation support for returning tasks.

We hope that the ideas and findings presented in the paper will offer a starting point for new research initiatives on patterns and temporal dynamics in recurrent user interests. Many aspects - such as individual differences between users - require future research, and it remains a challenge to develop tools to effectively support these recurrent activities.

6. ACKNOWLEDGEMENT

This research has been co-funded by the European Commission within the eContentplus targeted project OpenScout, grant ECP 2008 EDU 428016.

7. REFERENCES

- [1] F. Abel, E. Herder, G.-J. Houben, N. Henze, and D. Krause. Cross-system user modeling and personalization on the social web. *UMUAI - Journal of User Modeling and User-Adapted Interaction*, forthcoming.
- [2] E. Adar, J. Teevan, and S. T. Dumais. Large scale analysis of web revisitation patterns. In *CHI*, pages 1197–1206. ACM, 2008.
- [3] K. Bischoff, C. S. Firan, W. Nejdl, and R. Paiu. Can all tags be used for search? In *CIKM*, pages 193–202. ACM, 2008.
- [4] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30(1-7):107–117, 1998.
- [5] F. Carmagnola, F. Cena, L. Console, O. Cortassa, C. Gena, A. Goy, I. Torre, A. Toso, and F. Vernero. Tag-based user modeling for social multi-device adaptive guides. *User Model. User-Adapt. Interact.*, 18(5):497–538, 2008.
- [6] L. D. Catledge and J. E. Pitkow. Characterizing browsing strategies in the world-wide web. *Computer Networks and ISDN Systems*, 27(6):1065–1073, 1995.
- [7] F. Chierichetti, R. Kumar, and A. Tomkins. Stochastic models for tabbed browsing. In *WWW*, pages 241–250. ACM, 2010.
- [8] S. Goel, A. Z. Broder, E. Gabrilovich, and B. Pang. Anatomy of the long tail: ordinary people with extraordinary tastes. In *WSDM*, pages 201–210. ACM, 2010.
- [9] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Information retrieval in folksonomies: Search and ranking. In *ESWC*, pages 411–426, 2006.
- [10] R. Kumar and A. Tomkins. A characterization of online browsing behavior. In *WWW*, pages 561–570. ACM, 2010.
- [11] H. Obendorf, H. Weinreich, E. Herder, and M. Mayer. Web page revisitation revisited: implications of a long-term click-stream study of browser usage. In *CHI*, pages 597–606. ACM, 2007.
- [12] A. G. Parameswaran, G. Koutrika, B. Bercovitz, and H. Garcia-Molina. Recexplorer: recommendation algorithms based on precedence mining. In *SIGMOD*, pages 87–98, 2010.
- [13] M. Rasmussen and G. Karypis. gcluto - an interactive clustering, visualization, and analysis system. Technical report, Karypis Lab, 2004.
- [14] L. Tauscher and S. Greenberg. How people revisit web pages: empirical findings and implications for the design of history systems. *Int. J. Hum.-Comput. Stud.*, 47(1):97–137, 1997.
- [15] S. K. Tyler and J. Teevan. Large scale query log analysis of re-finding. In *WSDM*, pages 191–200. ACM, 2010.
- [16] M. van Setten, R. Brussee, H. van Vliet, L. Gazendam, Y. van Houten, and M. Veenstra. On the importance of “who tagged what”. In *Proceedings of the Workshop on the Social Navigation and Community based Adaptation Technologies at AH 2006*, pages 552–561, Dublin, Ireland, 2006.
- [17] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *WSDM*, pages 177–186. ACM, 2011.