Navigation Patterns from and to Social Media

Michael Gamon, Arnd Christian König

Microsoft Research
One Microsoft Way
Redmond, WA 98052
{mgamon,chrisko}@microsoft.com

Abstract

With the rapid rise of social media content, the question how people navigate to and from social media becomes important in order to understand what new tools and approaches are most beneficial to allow users to discover and interact with this social media data. In this study we analyze session log data to describe navigation patterns involving these sites. Our findings suggest that there are distinct patterns of insular navigation behavior around social media, with blogs being the most transitory in the way users interact with them.

Introduction

The volume of social media content on the web has increased dramatically over the past years (Technorati 2008 State of the Blogosphere Report). This in turn raises questions regarding discoverability and navigation: How do users navigate to social media? Where do they go from social media? Is there a lot of navigational cross-over from social media domains to other sources of information? How much of social media content is discovered via search engines? These are some of the basic questions we try to address in this paper. The data we analyze consists of session logs collected from Microsoft Live toolbar users across a range of dates in 2008. In previous research, such log data has been used for a variety of purposes (Agichtein et al. 2006, Poblete and Baeza-Yates 2008, Jones et al. 2006, White et al. 2007, amongst others). More general investigations of web user behavior include Sadagopan and Li 2008, White and Drucker 2007, Cockburn and McKenzie 2000, Adar et al. 2008, and Teevan et al. 2006. In this paper we provide an empirical analysis of user behavior with respect to social media sites, news sites, search, and other sites.

Data

We analyzed session log data collected through the Microsoft Live toolbar from 2008. We focused on data from the US-English market. Due to the size of the logs, we restricted our analysis to four dates from January through August 2008.

We need to emphasize from the start that the data we present should not be taken as representative for "internet users" in general. The users of a particular kind of mainstream toolbar, while numerous, are demographic subset of all internet users.

Each entry in the session logs consists of a series of URLs and timestamps per user session. Each session that we extracted belongs to a single unique toolbar installation, which in most cases will correspond to a single unique user.

Methods

Extracting Trails

We split sessions into smaller units, which we call *trails* following the convention in White and Drucker 2007. Trails are temporal sequences of visited URLs from a session where each visit does not exceed a given duration. A trail can thus be thought of as a temporal sequence of nodes (URLs) with transitions between them. In this temporally linear interpretation of a trail, "backtracking" or "branching" are not represented, reflecting our main interest in general domain transitions rather than individual user patterns (as in White and Drucker 2007).

We also make a distinction between *action* trails and *domain* trails. In an action trail, a node corresponds to a fully specified URL. Such a node corresponds to a single navigation action, for example entering a search query, following a link, reading a separate blog post page from the same blog. In contrast, a domain trail subsumes all navigation within a major domain under a single node.

Site Categorization

We categorize sites/URLs into one of the following categories:

1. Social sites: these are social media sites where the focus is on sharing media and navigating and building social networks and communities
2. Blog sites: dedicated sites for publishing blogs
3. Search: a search query performed on a web search engine

4. News: a news site
5. Other: all other sites

These categories are identified by matching against a list of known URLs, as shown in Table 1. These lists of URLs are based on manual categorization of the most frequently visited social sites and blog sites in our data. A web search is defined as a URL pointing to one of the major web search engine pages and containing a query. We did not take into account specialized blog search engines since they are used very rarely in our data.

Social sites	myspace, spaces.live, youtube, orkut, facebook, bebo, hi5, fotolog, friendster, metroflog, badoo, schuelervz, bilddagboken, wer-kennt-wen, tagged, skyrock, plentyoffish, studivz, mixi, netlog, wretch, hyves, dailymotion, piczo, myvideo, flickr, webshots, photobucket, shutterfly
Blogs	Typepad, LiveJournal, Multiply, Populum, Blogspot, Blogdumps, Bluehost, WordPress, Blogger, Typolis, Blogsome, TrendyFriendy, FusePress, Aeonity, Vox, Hipeople, BlogDrice, Steeky, Weblogs, Webmunism, NireBlog, Blog.ca, DABU, squarespace, iblogs, Eponym, BlogEasy, BlogFuse, Tabulas, Invisiblog, Memebot, Blogomonster, Blogspoint, Cool-blog, twoday.net, BlogCity, Opera, Xanga, netflog
News	List of 3500 national and international news sources.
Search	Google.com, Live.com, Yahoo search, Baidu.com, AOL search, Naver.com

Table 1: Site categorization.

Given the incompleteness of these lists, our approach only allows us to report a lower bound on the number of social site and blog visits.

To get an upper bound on categorization error, we manually analyzed a random sample of 500 URLs from the "other" category in the US data. This analysis showed that the "other" category is mostly made-up of general information, email, shopping and entertainment (including adult content). We did find 4% undetected news sites (often local news), 1.6% undetected search sites (minor search engines), 2% undetected social sites but no undetected blog sites. We also found 1.6% forum visits, which constitute a social medium but are not represented in our categorization.

Transition Probabilities as Markov Chains

Much previous research has used the simple but intuitive Markov Chain as the structure for modeling user behavior (Sadagopan and Li 2008, Sarukkai 2000 and Borges and Levene 1999). We can model basic category transition probabilities by using a set of states consisting of the 5 categories described in the previous section, plus a start symbol <START> and an end symbol <END> for the beginning and end of a session.

Results

We first report the percentage of visits and actions in each of the site categories. We then look at the total time spent in those categories. Next, we examine transition probabilities for individual actions and for changes from one category into another category to find out where people come from and go to when they navigate across categories. Similarly, we look at the most likely origin for a visit to a site category.

Where Do People Go? - Categories and Actions. The first analysis we performed relates to the overall distributions of both actions and visits to site categories. News sites account for 2.5% of actions and 3.5% of site visits. Social sites account for 12.5% of category visits, but for a much higher percentage (27.4%) of actions. The reverse holds for uncategorized "other" sites: those account for 74.5% of all visits, but only for 64.3% of all actions. Search accounts for a relatively small amount of actions (5.6%) and visits (9.2%), and blogs only for less than one percent of actions or visits.

Where do People Spend Their Time? One might assume that the answer to this question closely parallels the results from the previous section; after all it is reasonable to assume that the number of actions performed in a site category determines the amount of time spent there. We found, however, that the time spent per action varies with the category of the site that the action is performed in. Specifically, there is a pronounced shift towards spending a larger proportion of time on sites in the uncategorized "other" class compared to the action distribution. Social sites garner a smaller fraction of the total dwell time compared to the fraction of actions they account for. Both news and blogs, on the other hand show larger total dwell times than their share of actions would predict.

What do people do next? - Action transitions. Action transition probabilities from each of the site categories are shown in the "action" columns in Table 3. The site category that the action comes from is shown in the columns; rows correspond to the site categories that the action leads to. For example, the probability of an action leading from Search to Blog, is located in the Search column and the Blog row. This data illustrates the tendency of a user to "stay" within a site category. The highest probability in the set of transitions from each site category is boldfaced in the table.

Blogs: Within blog sites, the combined probability of performing another action within the blog site category or ending a session is above 75%. When an action takes the user outside of the blog site category, an action on an uncategorized site is the most likely candidate at 14.1%, followed by a search action at 7.7% and news with only 0.7%.

Social sites: Actions on social sites are very likely to be followed by other actions on social sites, or to end a session, with a combined probability of more than 98%. Amongst all site categories, social sites have by far the strongest tendency to retain the user within this category.

News: Users performing a navigation action on a news site again are most likely to stay within the news category. This tendency is much weaker than the one observed for social sites, but stronger than the one for blog sites. When

transferring to another site category, the uncategorized "other" class is the most preferred target.

Search: When performing a search, the user is most likely to either follow up with an additional search, or navigate to an uncategorized site. The former seems to indicate that query reformulation as a strategy to obtain better search results is common. Search rarely ever leads to a blog site with a likelihood of only 0.3%.

Overall, the results here show a clear tendency to stay within a site category rather than to go across categories. Social sites are the most insular when it comes to user actions: these sites fulfill a specific purpose and as such the user clearly has a strong tendency to perform multiple actions within the same site category. News and blogs, on the other hand, serve the purpose of providing information, and in the case of blogs, personal commentary. As such, it is natural that the tendency of the user to stay within this site category with their next action is less pronounced than with social sites. Interestingly - with the exception of search - blogs have the least tendency to be "insular", constituting a relatively "transitory" category.

Where do people go next? - Category transitions.

The "categ" columns in Table 2 show the probabilities of transition from one category to another. For these probabilities, only a category change is counted as a transition, so a 0.25 probability of change from category A to category B means that if there is a change from category A to another category, the probability of that change going into category B is 0.25.

Blogs: When users leave a blog site, they tend to navigate to an uncategorized site (26%). Actions on a search site or another blog site are the next most likely candidates. Navigation to a social site is relatively unlikely (2%). Users are likely to transition to another blog site, presumably due to the easy availability of blog-links through blogrolls.

Social sites: If a user exits a social site, the most likely target is an uncategorized "other" site. Search is the next likely choice (4.6% likelihood). Another social site can be the target at 3.9%, news domains have less than 1% probability of being the next target. Ending a session from a social site is much more likely than for any other category.

News: With a 39.5% probability, transitioning out of a news site will lead to an uncategorized site. A search site or another news site are the next likely candidates, with roughly similar probability.

Search: Search is most likely to lead to an uncategorized site (70.7%). The likelihood to change from one search site to another is the second most likely choice, at 3.3%, probably a consequence of a change of search engines to improve results. Social sites and news account for about 2.5% each.

Where do people come from? - Inbound probabilities. We collected the probabilities of a previous site category visit leading to the current site category. A probability of 0.25 from A to B in this case means that there is a 25% probability that a user who is visiting site category B has arrived at B from site category A. For example, the probability of arriving at a Blog site from a Search site can be found in the row labeled Blog and the column labeled Search.

The first column shows the probability that a user starts their session at a site category. This allows to draw a distinction between two behaviors: (a) coming across a site during navigation versus (b) targeting a specific site with a specific intent from the get-go. The fact that social sites have a 90% probability to be visited at the beginning of a session clearly indicates that they are serving a single and specific purpose. Similarly, news sites tend to occur at the beginning of a session, with the user having a specific information need and a specific site they consult for that need. Blogs, on the other hand, have a much lower probability to be visited at the start of a session. This observation could indicate that blogs are often visited as a secondary source of information, with various entry points through other sites. Blog visits also tend to end a session with less probability than for example news visits, another indication of their "transitory" nature in navigation.

Note also that 16% of blog visits originate from search while only 0.3% of search actions lead to a blog visit.

In order to gain some better understanding of the transitory character of blogs, we manually examined a sample of the non-blog user actions leading to and from blogs. About one fourth of all these actions were through images on a blog site indexed through an image search engine. Besides search, another big contributor to blog visits was what could be defined as an information-browsing session: the user navigates from a site that provides information about a topic to a blog site, presumably through a link provided on the first site. This accounted for about a third of the non-blog actions leading to blogs. Leaving a blog site also leads to other information-providing sites or the location of an image or video with roughly the same probability.

We also performed a more detailed manual breakdown of traffic to blogs and an in-depth analysis of search queries that lead to blog posts see Gamon and König (to appear).

Search Queries Leading to Blogs

Are blogs relevant when surfaced by search? In the absence of click-through data or human relevance judgments, we examined two measures for search-to-blog transitions that can be seen as an indicator of the relevance of a blog search result:

(a) The probability of the trail ending at the clicked search result

From	<START>		Blog		Social Site		News		Search		Other	
To	action	categ	action	categ	action	categ	action	categ	action	categ	action	categ
Blog	0.002	0.002	**0.522**	0.130	0.0001	0.002	0.0005	0.002	0.003	0.007	0.0004	0.002
Other	**0.731**	**0.730**	0.141	0.264	0.013	0.190	0.154	0.395	0.417	**0.707**	**0.757**	0.376
Search	0.066	0.066	0.077	0.162	0.004	0.046	0.027	0.078	**0.418**	0.033	0.037	0.113
Social Site	0.170	0.170	0.020	0.021	**0.864**	0.039	0.008	0.009	0.037	0.024	0.012	0.017
News	0.032	0.032	0.007	0.014	0.0004	0.005	**0.585**	0.062	0.014	0.025	0.005	0.019
<END>			0.233	**0.407**	0.118	**0.718**	0.226	**0.454**	0.111	0.204	0.187	**0.473**

Table 2: Action and category transition probabilities

From	<START>	Blog	Social Site	News	Search	Other
To						
Blog	0.389	0.115	0.029	0.012	0.161	0.295
Other	0.670	0.001	0.013	0.013	0.064	0.239
Search	0.420	0.003	0.017	0.016	0.019	0.477
Social Site	0.907	0.001	0.018	0.002	0.015	0.075
News	0.641	0.001	0.007	0.045	0.049	0.257

Table 3: Inbound probabilities for site category transitions.

(b) the average dwell time at a blog before the next transition.

The first measure indicates that the user may have attained their information goal. A long dwell time on the blog is a plausible indicator for having reached a site with meaningful and relevant information. We compared the results from these two measures on search-to-blog and search-to-non-blog transitions in our trail data. We found that the probability of a trail ending was significantly higher for blog than for other search results (28.7% vs. 23.51%). The average dwell time was also significantly longer (1 Minute, 33 seconds vs. 1 Minute, 7 seconds).

We believe that this can serve as evidence that blog results are likely to be relevant as a search result.

Conclusion

A question that our initial exploration leads to is: Is there a need for improved tools and services to navigate and search social media within the mainstream search engines and portals? We do believe this to be the case. The main reasons for this hypothesis are:

1. Blog content is large and rapidly growing
2. We find little usage of specialized blog search engines
3. Only a small fraction of searches lead to blog content
4. A sizeable portion of blog visits come from search
5. When blogs are reached through search, dwell time indicates that the blogs are relevant

For a more in-depth analysis of the data, see Gamon and König (to appear).

References

E. Adar, J. Teevan and S. Dumais. 2008. Large Scale Analysis of Web Revisitation Patterns. CHI, pages 1197-1206..

E. Agichtein, E. Brill, and S.T. Dumais. 2006. Improving web search ranking by incorporating user behavior information. In SIGIR, pages 19-26.

J. Borges and M. Levene. 1999. Data Mining of User Navigation Patterns. Workshop on Web Usage Analysis and User Profiling, pages 31-36.

A. Cockburn and B. McKenzie. 2001. What Do Web Users Do? An Empirical Analysis of Web Use. International Journal of Human-Computer Studies, pages 903-922.

M. Gamon and A. C. König. To appear. Exploring Navigation Patterns around Social Media in User Session Data. Microsoft Technical Report.

R. Jones, B. Rev, O. Madani, and W. Greiner. 2006. Generating query substitutions. In WWW, pages 387-396.

The Pew Internet and American Life Project. Pew Research Center. www.pewinternet.org.

B. Poblete and R. Baeza-Yates. 2008. Query-sets: using implicit feedback and query patterns to organize web documents. WWW, pages 41-50.

N. Sadagopan and J. Li. 2008. Characterizing typical and atypical user sessions in clickstreams. Proceedings of WWW, pages 885-893.

R. R. Sarukkai. 2000. Link Prediction and Path Analysis Using Markov Chains. Proceedings of WWW, pages 377-386.

Technorati. State of the Live Web, April 2007. technorati.com/weblog/2007/04/328.html.

Technorati. State of the blogosphere 2008. technorati.com/blogging/state-of-the-blogosphere/.

J. Teevan, E. Adar, R. Jones and M. Potts. 2006. History Repeats Itself: Repeat Queries in Yahoo's Logs. Proceedings of SIGIR 2006, pages 703-704.

R. W. White and S. M. Drucker. 2007. Investigating behavioral variability in web search. Proceedings of WWW, pages 21-30.

R.W. White, M. Bilenko, and S. Cucerzan. 2007. Studying the use of popular destinations to enhance web search interaction. Proceedings of SIGIR, pages 159-166.