How Web 1.0 fails: the mismatch between hyperlinks and clickstreams

Lingfei Wu · Robert Ackland

Received: 10 February 2014 / Revised: 24 April 2014 / Accepted: 13 May 2014 / Published online: 10 June 2014
© Springer-Verlag Wien 2014

Abstract The core of the Web is a hyperlink navigation system collaboratively set up by webmasters to help users find desired information. While it is well known that search engines are important for navigation, the extent to which search has led to a mismatch between hyperlinks and the pathways that users actually take has not been quantified. By applying network science to publicly available hyperlink and clickstream data for approximately 1,000 of the top Web sites, we show that the mismatch between hyperlinks and clickstreams is indeed substantial. We demonstrate that this mismatch has arisen because webmasters attempt to build a global virtual world without geographical or cultural boundaries, but users in fact prefer to navigate within more fragmented, language-based groups of Web sites. We call this type of behavior “preferential navigation” and find that it is driven by “local” search engines.

Keywords Clickstream · Hyperlink · Search engine · Navigation · Social network analysis

1 Introduction

Invented by Tim Berners-Lee in 1991, the World Wide Web is regarded as the “largest human information construct in history” (http://webscience.org/web-science/web-science-home/). The Web is commonly understood to have had three overlapping phases of development or eras. Under Web 1.0, webmasters provide content that is consumed by users, while Web 2.0 blurs the distinction between webmasters and users, with blogging tools, social network sites (e.g., Facebook), and microblog services (e.g., Twitter) enabling non-technical people to both produce and consume content (“prosumption”) (Ritzer and Jurgenson 2010). Web 3.0, or the Semantic Web, involves technologies that make the Web more machine-readable, leading to a “Web of data,” which is an evolution of the Web 1.0 “Web of documents” (Shadbolt et al. 2006).

A common feature of all three phases is the use of technologies to help people find Web content. With Web 1.0, and to a lesser extent Web 2.0, the core enabling technology is the hyperlink, which allows users to efficiently move around the Web, while Web 3.0 envisages automated agents finding content on behalf of users by drawing on users’ browsing habits.

Here, we use novel data and methods to investigate the extent to which Web navigation is based on hyperlinks. We construct clickstream and hyperlink networks comprised of the same 980 Web sites. In the clickstream network, a directed weighted edge between Web sites $i$ and $j$ indicates the percentage of global Web users who visited Web site $i$ and then immediately visited Web site $j$. The clickstream network thus shows the pathways that people are taking as they navigate the Web. In the hyperlink network, a directed unweighted edge between Web sites $i$ and $j$ indicates that $i$ hyperlinks to $j$, and hence the hyperlink network shows the pathways that webmasters are creating for users.

Our analysis reveals a substantial mismatch between the hyperlink and clickstream networks, allowing us to conclude that in navigating the Web, users tend to create their
own pathways rather than following hyperlinks laid out by webmasters. This mismatch between hyperlinks and clickstreams reveals different preferences of webmasters and users: while webmasters work collaboratively to build a fully connected online society, users in fact only navigate within the fragmental parts of the Web that they favor, a behavior which we term “preferential navigation.” We conclude by suggesting a simple approach that can be used by a webmaster to quantify the hyperlink–clickstream match or overlap for a given Web site, providing actionable information that can be used to improve this match.

2 From hyperlink analysis to clickstream studies

The importance of hyperlinks to the Web has led to a large amount of research, with applied physics research into power law properties of degree distribution in hyperlink networks and models that explain their emergence (Barrat and Albert 1999), and computer scientists examining how hyperlink structures can be used to improve Web search (Kleinberg 1999; Page et al. 1997). Social scientists have contributed studies on the motivations for hyperlink creation (Park et al. 2004; Shumate and Dewitt 2008; Ackland and O’Neil 2011), the political implications of power laws on the Web (Hindman et al. 2003), and the extent to which structural relations between countries (arising from economics, culture, language, and geography) are reflected in hyperlink networks (Barnett et al. 2001; Barnett and Sung 2005; Park et al. 2011).

However, as noted in recent research on e-learning systems, hyperlinks are too simple to represent the rich connections between documents that are created by users’ various online activities (Zhuge 2009, 2011). While both hyperlinks and clickstreams can be used to show connections between documents, the former reflects the preferences of a relatively small number of webmasters, whereas the latter is the collaborative product of potentially massive numbers of users. Furthermore, the use of search engines, bookmarks, default home pages, and historical viewing records means there are many ways through which users can move between Web sites when there is no hyperlink connection (Qiu et al. 2005; Meiss et al. 2010).

Although clickstreams have been analyzed since the early days of the Web (Catledge and Pitkow 1995), it is only in the past decade that they have been studied from a network perspective. Much of this work has been focused on intra-Web site clickstream networks in order to improve the design of sites, including social networks (Schneider et al. 2009), tagging systems (Cattuto et al. 2007), news sharing systems (Wu and Huberman 2007), and citation systems (Bollen et al. 2009). Research into recommendation systems and mobile computation is also making use of clickstream network analysis (Kim et al. 2005; Yamakami 2006), and professional software for clickstream network analysis (Brainerd and Becker 2001) has also been developed.

Researchers have also studied inter-Web site clickstream flows (Qiu et al. 2005; Meiss et al. 2010), and this is the focus of the present study. However, a key difference between the present and earlier studies of inter-Web site clickstreams is that we jointly analyze the hyperlink and clickstream networks, allowing us to quantify the extent of the mismatch between these networks and identify underlying causes. While Barnett and Park 2005 also compared the structure of hyperlink and clickstream networks, they used aggregated country-level network data, while the present study is focused on hyperlink and clickstream connections between individual Web sites.

3 Tools for collecting hyperlink and clickstream data

We selected the top 1,000 Web sites according to Google’s traffic statistics in November 2010 (http://www.google.com/adplanner/static/top1000/). We then used Alexa (http://www.alexa.com) to retrieve the daily traffic to these Web sites (which is averaged over three months) and also the daily clickstreams between them. The sum of clickstreams to a given site will be less than or equal to the traffic to that site, since clickstreams only refer to visits from the set of 1,000 Web sites, while traffic is all visits to the site. According our calculation based on Google’s statistics, these 1,000 Web sites account for more than 96% of global Web traffic during the period of the data collection. The clickstream network contains 12,008 directed and weighted edges, where an edge between Web sites $i$ and $j$ indicates the percentage of global Web users who visited $j$ immediately after visiting $i$.

We then used VOSON (http://voson.anu.edu.au/), which is software for hyperlink network construction and analysis created by one of the authors (Ackland 2010), to construct a hyperlink network where a directed and unweighted edge between Web sites $i$ and $j$ indicates that $i$ contains a hyperlink to $j$. The hyperlink network contains 15,907 edges.

Twenty sites were dropped due to a lack of data, and thus, our analysis is for the remaining 980 sites. Further, as Alexa reports a maximum of ten largest inbound and outbound clickstreams for each Web site, and the version of the VOSON Web crawler that was used only collected a maximum of 1,000 outbound hyperlinks for each site, the two constructed networks necessarily do not include all of the clickstreams and hyperlinks between these 980 sites. It is important to note that our data do not allow us to know exactly how a person navigates from Web site $i$ to $j$: navigation may occur either through a hyperlink, a search
engine, or the user typing the URL into the browser (or equivalently, following a bookmark).

4 The mismatch between hyperlinks and clickstreams

The hyperlink network and the clickstream network are shown in Fig. 1, with edges that are common to both networks drawn in red. The fact that the clickstream and hyperlink networks are comprised of the same nodes allows us to check for overlaps between the two sets of directed edges. Only 2,580 out of the 15,907 hyperlinks overlap with the 12,008 clickstreams, meaning that a large proportion of hyperlinks are “useless” in the sense that they connect sites that did not exchange any traffic during the data collection period. The clickstreams transported by hyperlinks only account for 33% of the sum of all clickstreams. This percentage gives an upper bound of the hyperlink-moderated clickstreams, since we collect more hyperlinks than clickstreams for each Web site.

In other words, the actual proportion clickstreams driven by hyperlinks are likely to be even smaller. In Table 1 we compare the hyperlink and clickstream networks in terms of several structural properties (Freeman 1979; Watts and Strogatz 1998). We note that the Web sites tend to cluster into small, local groups in the clickstream network (which has smaller global transitivity and density but is larger in average local transitivity). We conjecture that this structural difference reflects different navigation behaviors underlying the two networks; we explore this further below.

4.1 Different preferences of webmasters and users

To investigate the reason for the observed mismatch between the hyperlink and clickstream networks, we simulated user navigation in both networks using a label propagation algorithm (Raghavan et al. 2007), which works as follows. Each Web site is initially assigned a unique label and then at every step of the simulation, each Web site adopts the most popular label in its neighborhood. The process continues until there is no further change in labels, and Web sites with the same label are clustered into a community.

To illustrate, Web site 2 in the example network shown in Fig. 2 is visited by users from three Web sites, 0, 1, and 6. After several steps of simulation, Web site 6 is “occupied” by users coming from Web site 1, therefore 2 will adopt the same label as 6 and 1. Eventually, Web sites 1, 6, and 2 are clustered into a community, with Web site 0 belonging to another. The label propagation algorithm thus allows us to simulate a group of users who share similar interests diffusing on the network and labeling all visited Web sites until coming to a Web site that has been “occupied” by another group of users. The detected communities therefore correspond to the preferences of different users.

The simulation identified six communities from the clickstream network (Fig. 3), which broadly coincided with the visual clustering provided by the Fruchterman–Reingold algorithm. Web sites in each community

![Fig. 1 Hyperlink network (a) and the clickstream network (b). The size of nodes denotes the logarithm of traffic. Edges that are common to both networks are drawn in red. The node layout method is Fruchterman and Reingold (1991), and for ease of comparison, both the two networks have the node layout that was computed for the clickstream network (color figure online).](image)

![Table 1 Network statistics for the hyperlink and clickstream networks](table)
generally share the same language (with an accuracy rate of 96%, validated by human coding), and we therefore label the communities according to these languages: Polish community, Korean community, Russian community, Japanese community, Chinese community, and Euro-American community (over 85% of Web sites within this community are in English). As indicated by Table 2, the Euro-American community is the largest, accounting for 645 sites (66% of the total) and 7,695 links (64% of the total), while the smallest (Polish) consists of only four Web sites.

We conducted the simulation on the hyperlink network, but found that all Web sites were clustered into a single community, providing compelling evidence that webmasters and users exhibit very different preferences. The hyperlink network is constructed by webmasters, who link their Web sites to those Web sites they think visitors will be interested in. If users followed the pathways setup by webmasters, there would be a very high level of diffusion, as seen in the simulation conducted on the hyperlink network. However, the communities identified in the clickstream network suggest that the linking structure setup by the webmasters does not meet the requirements of users, who in fact prefer to navigate within local, language-based fragments of the Web. We call this behavior “preferential navigation”.

4.2 Preferential navigation driven by local search engines

It was mentioned earlier that our clickstream data do not allow us to know the exact process by which a user visits two Web sites \( i \) and \( j \) in succession. There are three possibilities: (1) follow a hyperlink from \( i \) to \( j \); (2) enter the URL for \( j \) directly into the browser after visiting \( i \) (or equivalently, follow a bookmark for \( j \)); or (3) navigate from \( i \) to a search engine, and then navigate to \( j \) by clicking on a search result.

We have already shown that (1) does not appear to account for a large proportion of clickstreams. The question we now pose is: how important are search engines in user navigation, and what is their role in enabling preferential navigation? To answer this, we examined the role of search engines in the clickstream network. Firstly, we divided the clickstream network into sub-networks on the basis of language (this was done manually in light of the fact that the label propagation algorithm had a 4% error rate). For each sub-network (the Polish community was excluded due to its small size), we calculated three centrality measures: degree, betweenness, and closeness. According to Freeman (1979), these quantities reflect the importance of a node in different aspects: degree indicates the activity of a node, betweenness is a measure of a node’s ability to control the flow in the network, and closeness shows a node’s efficiency in resource transmission. We find that the five search engines, google.com (Euro-American), baidu.com (Chinese), yahoo.co.jp (Japanese), yandex.ru (Russian), and naver.com (Korean), have the highest values in terms of all three centrality measures. This finding clearly points to the predominant role of search engines in facilitating navigation.
The above analysis only reveals the static, structural importance of search engines in the clickstream network; to further investigate the role of search engines in navigation, we compared the clickstreams driven by these five search engines with the clickstreams moderated by hyperlinks. The five search engines moderate over 42% of total clickstreams, with google.com accounting for over a half of the clickstreams moderated by these search engines, followed by baidu.com, yahoo.co.jp, yandex.ru, and naver.com. As mentioned above, hyperlinks only moderate 33% of all the clickstreams. Thus, it is reasonable to conclude that users rely more on search engines than hyperlinks in surfing the Web.

So how exactly do these search engines drive clickstreams? To address this question, we introduce a novel approach for analyzing clickstreams called “popular pathway analysis.” This approach is inspired by the “maximizing chain strengths analysis” used in studies of food webs (Garlaschelli et al. 2003). In each of the five communities, we start from a Web site ranking first in traffic according to Alexa’s statistics. Then, in each of the following steps, we choose the strongest (with the largest weight) outbound clickstream. We stop at the fourth step and draw the clickstream sub-network comprising the Web sites included in the selected clickstreams, which shows the “most likely pathways” a typical user in the communities may take (Fig. 4).

We found very similar circulations of clickstream in the five communities examined. In these circulations, users start from a “local” search engine and return to it repeatedly after visiting other Web sites. By using the term “local,” we mean that a search engine is only popular within a group of users speaking the same language. As search engines usually take into account the feedback (clicks) of users in ranking documents, a search engine frequently visited by users using the same language is more likely to recommend Web pages in this language (e.g., baidu.com usually shows Chinese Web pages on the first

Fig. 3 Six language-based communities were detected from the clickstream network. The Web sites in different communities are shown in different colors. The nodes that are assigned to an incorrect community by the label propagation algorithm are plotted in white.
Fig. 4 “Popular pathways” in different communities. The squares show the search engines as starting points, and the circles denote other Web sites on the pathways. The weights on clickstreams indicate traffic measured in millions of unique users. The pathway colors correspond to the colors of communities in Fig. 3

page), which in turn reinforces the preferences of users in choosing search engines.

5 Conclusion

Our novel approach for quantifying the mismatch between clickstreams and hyperlinks provides substantial evidence that users create their own pathways on the Web instead of following hyperlinks passively. We contend that this mismatch originates from the different preferences of webmaster and user: the former setup links to connect to each other’s Web site collaboratively, leading to a highly connected hyperlink network, while the latter use local search engines to guide preferential navigation, which eventually results in a more fragmented, clustered network. It should be noted that although we focus on language in the current study, there are other factors (e.g., culture and commerce) that shape clickstream flows. In future work, we intend to explore these other factors more fully by using regression to identify outliers in the clickstream network. In a more fragmented, clustered network. It should be noted that although we focus on language in the current study, there are other factors (e.g., culture and commerce) that shape clickstream flows. In future work, we intend to explore these other factors more fully by using regression to identify outliers in the clickstream network.

The findings in this study are relevant to several areas of Web development. For example, hyperlink-based ranking algorithms may provide biased estimates of Web site relevance (Page et al. 1997; Kleinberg 1999) since they are derived from the hyperlink network structure created by webmasters, and our research has shown that Web users do not tend to follow hyperlinks. Our findings on the extent of preferential navigation and the underlying causes are relevant to search engine companies who aim to become successful in more than one language-based community.

While we provided above aggregate measures of the mismatch between clickstreams and hyperlinks for a collection of websites, our new approach can also be used to benchmark and monitor particular Web sites. For example, we can define for a given Web site i the \( H_i \) index of hyperlinks and clickstreams as \( \frac{|H_i \cap C_i|}{|H_i| + |C_i| - |H_i \cap C_i|} \), where \( H_i \) is the set of sites which i hyperlinks to and \( C_i \) is the set of sites which i directs clickstreams toward. The \( H_i \) index is analogously defined. The hyperlink–clickstream match indexes can be built into hyperlink collection and analysis tools to provide Web site owners with actionable information so that the hyperlink pathways to and from their Web sites become better aligned with preferential navigation.

We predict that as the Web becomes increasingly intelligent, the hyperlink structure will gradually adapt to the clickstream structure, leading to a decrease in the mismatch. In Web 1.0, webmasters are the major constructors of the hyperlinks, and their limited information on user preferences leads to the mismatch. In the era of Web 2.0, users are able to hyperlink by themselves, for example linking from Facebook homepages to blogs. By encouraging users to contribute to the content of websites, webmasters incorporate user preference in setting up hyperlinks, and hence decrease the level of the mismatch.
that are more consistent with user Web surfing preferences. 

appropriate Web sites, leading to the creation of hyperlinks will continue to decrease, as the Web will be able to analyze users’ historical surfing records and recommend appropriate Web sites, leading to the creation of hyperlinks that are more consistent with user Web surfing preferences.

Acknowledgments We thank Jonathan J. H. Zhu, Lexing Xie, Paul Thomas, Hai Liang, and the reviewers for providing comments on an earlier version of this paper.

References

Ackland R (2010) WWW hyperlink networks. In: Hansen DL, Shneiderman B, Smith MA (eds) Analyzing social media networks with NodeXL: insights from a connected world. Morgan-Kaufmann, Burlington

Ackland R, O’Neil M (2011) Online collective identity: the case of the environmental movement. Soc Netw 33:177–190

Barabási A, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509

Barnett GA, Chon BS, Rosen D (2001) The structure of internet flows in cyberspace. Netw Commun Stud (NETCOM) 15(1–2):61–80

Barnett GA, Park HW (2005) The structure of international internet hyperlinks and bilateral bandwidth. Annales des Telecommunications 60(9–10):1115–1132

Barnett GA, Sung EJ (2005) Culture and the structure of the international hyperlink network. J Comput Med Commun 11(1):217–238

Bollen J, Van de Sompel H, Hagberg A, Bettencourt L, Chute R, Rodriguez M, Balakireva L (2009) Clickstream data yields high-resolution maps of science. PLoS One 4(3):e4803

Brainerd J, Becker B (2001) Case study: e-commerce clickstream visualization. In: Proceedings of the IEEE symposium on information visualization 2001 (INFOVIS’01), p 153. IEEE Computer Society

Catledge LD, Pitkow JE (1995) Characterizing browsing strategies in the World Wide Web. Comput Netw ISDN Syst 27:1065–1073

Cattuto C, Loreto V, Pietronero L (2007) Semiotic dynamics and collaborative tagging. Proc Natl Acad Sci 104(5):1461

Freeman L (1979) Centrality in social networks conceptual clarification. Soc Netw 1(3):215–239

Fruchterman T, Reingold E (1991) Graph drawing by force-directed placement. Softw Pract Exp 21(11):1129–1164

Garlaschelli D, Caldarelli G, Pietronero L (2003) Universal scaling relations in food webs. Nature 423(6936):165–168

Hindman M, Tsioutsiouliklis K, Johnson J (2003) Googlesearch: how a few heavily-linked sites dominate politics on the web. In: Annual meeting of the midwest political science association, vol 4. Citeseer, pp 1–33

Kim D, Im I, Atluri V (2005) A clickstream-based collaborative filtering recommendation model for e-commerce. In: Seventh IEEE international conference on e-commerce technology, 2005 (CEC 2005). IEEE, pp 84–91

Kim DH, Atluri V, Bieber M, Adam N, Yesha Y (2004) A clickstream-based collaborative filtering personalization model: towards a better performance. In: Proceedings of the 6th annual ACM international workshop on web information and data management. Association for Computing Machinery

Kleinberg J (1999) Authoritative sources in a hyperlinked environment. J ACM (JACM) 46(5):604–632

Meiss MR, Gonalves B, Ramasco JJ, Flammini A, Menczer F (2010) Agents, bookmarks and clicks: a topical model of web navigation. In: Proceedings of the 21st ACM conference on Hypertext and hypermedia

Page L, Brin S, Motwani R., Winograd T (1997) Pagerank: bringing order to the web. www.pcd.stanford.edu/~page/papers/pagerank. Accessed 29 Jan 2001

Park HW, Barnett GA, Chung CJ (2011) Structural changes in the 2003–2009 global hyperlink network. Glob Netw 11(4):522–542

Park HW, Kim CS, Barnett GA (2004) Socio-communicational structure among political actors on the web in South Korea: the dynamics of digital presence in cyberspace. New Media Soc 6(3):403–423

Qiu F, Liu Z, Cho J (2005) Analysis of user web traffic with a focus on search activities. In: Proceedings of the international workshop on the web and databases

Raghavan UN, Albert R, Kumara S (2007) Near linear time algorithm to detect community structures in large-scale networks. Phys Rev E 76(3):036106

Ritzer G, Jurgenson N (2010) Production, consumption, prosumption. J Consum Cult 10(1):13

Schneider F, Feldmann A, Krishnamurthy B, Willinger W (2009) Understanding online social network usage from a network perspective. In: Proceedings of the 9th ACM SIGCOMM conference on internet measurement conference. ACM, pp 35–48

Shadbolt N, Hall W, Berners-Lee T (2006) The semantic web revisited. Intell Syst IEEE 21(3):96–101

Shumate M, Dewitt L (2008) The North/South divide in NGO hyperlink networks. J Comput Med Commun 13:405–428

Watts DJ, Strogatz SH (1998) Collective dynamics of small-world networks. Nature 393(6684):440

Yamakami T (2006) Regularity analysis using time slot counting in the mobile clickstream. In: Database and expert systems applications, 2006. DEXA’06. 17th international workshop on, IEEE, pp 55–59

Zhuge H (2009) Communities and emerging semantics in semantic hyperlink networks. J Comput Med Commun 13:405–428

Zhuge H (2011) Semantic linking through spaces for cyber-physical-socio intelligence: a methodology. Artif Intell 175(5–6):988–1019