Search Result Presentation: Supporting Post-Search Navigation by Integration of Taxonomy Data

Matthias Keller
Steinbuch Centre for Computing
Karlsruhe Institute of Technology
D-76128 Karlsruhe, Germany
matthias.keller@kit.edu

Patrick Mühlschlegel
Steinbuch Centre for Computing
Karlsruhe Institute of Technology
D-76128 Karlsruhe, Germany
uzbto@student.kit.edu

Hannes Hartenstein
Steinbuch Centre for Computing
Karlsruhe Institute of Technology
D-76128 Karlsruhe, Germany
hannes.hartenstein@kit.edu

ABSTRACT

As a result of additional semantic annotations and novel mining methods, Web site taxonomies are more and more available to machines, including search engines. Recent research shows that after a search result is clicked, users often continue navigating on the destination site because in many cases a single document cannot satisfy the information need. The role Web site taxonomies play in this post-search navigation phase has not yet been researched. In this paper we analyze in an empirical study of three highly-frequented Web sites how Web site taxonomies influence the next browsing steps of users arriving from a search engine. The study reveals that users not randomly explore the destination site, but proceed to the direct child nodes of the landing page with significantly higher frequency compared to the other linked pages. We conclude that the common post-search navigation strategy in taxonomies is to descend towards more specific results. The study has interesting implications for the presentation of search results. Current search engines focus on summarizing the linked document only. In doing so, search engines ignore the fact the linked documents are in many cases just the starting point for further navigation. Based on the observed post-search navigation strategy, we propose to include information about child nodes of linked documents in the presentation of search results. Users would benefit by saving clicks, because they could not only estimate whether the linked document provides useful information, but also whether post-search navigation is promising.

Categories and Subject Descriptors

H.5.4 [**Information Systems**]: Information Interfaces and Presentation – *Hypertext/Hypermedia;*

Keywords

Search Result Presentation, Taxonomies, Clickstreams

1. INTRODUCTION

Without systems for crawling and indexing Web contents, the largest information source of the planet could not be utilized as it is today. Consequently, search engines and related technologies have attracted a lot of research interest in recent years. Current search engines are document-centric in a way that they return a list of ranked documents. However, in their original context the documents are usually part of a site. Each site has its own, handcrafted information architecture. Information architecture, which is independent from the underlying technical system, is the way information is organized, labeled and linked. It is handcrafted, because humans know best how to organize information for human access. For example, information architects spend much effort on dividing the content in different sections, on finding meaningful labels and on arranging them in taxonomies for creating hierarchical menus. At the same time, the information architecture is of crucial importance for the usability and, thus, the success of the site. Today and in the near future automated keyword-extraction and clustering algorithms will not be able to solve this task as well as humans do. But developing the information architecture is not only human creativity but engineering. Methods as Card Sorting [1] are used to arrange pages based on surveys. User reactions are observed in usability labs with methods such as eye tracking. Web analytics are applied to iteratively increase user satisfaction and conversion rates [2]. Usability involves many factors, but as we outline in Section 3, well-designed Web site taxonomies, understood as the hierarchical arrangement of documents in different menu levels, are of particular importance.

Given the effort usability experts spend on designing taxonomies, one would expect that this semantic information is very valuable for augmenting Web search results. Surprisingly, except displaying extracted breadcrumb trails (cf. Section 2), current search engines do not yet utilize taxonomy data. The simple reason is that Web site taxonomies were not available in the past. HTML allows to model nested lists but not to define site-wide taxonomies for navigation. Instead, the taxonomies are only visually encoded. By this, we mean that humans can easily distinguish e.g. the main menu that represents the main content sections and thus the first level of the Web site taxonomy as well as the second navigation level based on its visual features and position. Machines in contrast cannot extract these semantics from the HTML code. However, two current developments change this situation:

- More and more Web sites use structured data markup to encode machine-readable knowledge. That includes, e.g., a semantic formalization of breadcrumb trails[1] specified by Schema.org, which is an initiative of the major search engine operators. Breadcrumb trails show the position of a page in a Web site taxonomy. By combining the breadcrumb trails of each page the whole Web site taxonomy can be retrieved.
- Novel data mining methods allow extracting taxonomy data from sites based on mining navigation elements. The search engine Google, e.g., is able to recognize breadcrumb trails in many cases, even if they do not have semantic annotations.

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

[1] Cf. http://ui-patterns.com/patterns/Breadcrumbs

Our own work includes a method for mining Web site menus [3] as foundation for extracting the underlying taxonomies.

In this paper we present a novel way of utilizing the newly available Web site taxonomies for Web search. In more detail, the contributions and the structure of this paper are:

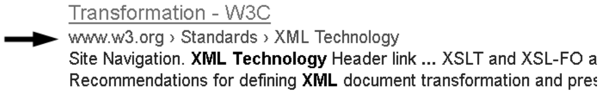
- In Section 3 we provide a definition for Web site taxonomies and analyze that usability experts attach great importance to them.
- Current search engines focus on delivering the most interesting documents, but we argue in Section 4 that this paradigm does not match current research results on search strategies. We conclude that search engines should also try to give information if a document is a suitable starting point for further navigation on the site. We provide illustrative examples that demonstrate how users can benefit from an enhanced presentation of search results as proposed in this paper.
- In Section 5, we present an empirical study based on usage data of three highly-frequented Web sites that supports our findings. The role taxonomies play in post-search navigation has not yet been researched. The study reveals that users tend to navigate along the edges of the Web site taxonomy down towards more specific results, when arriving from a search engine. This means that the user's next browsing steps can be predicted effectively and, moreover, that information about the most interesting next browsing steps can be included in the presentation of the search results.

2. RELATED WORK

The PageRank algorithm [4], which is the foundation of Google's ranking method, models the Web as a graph, defined by documents and hyperlinks. That model is an abstraction of the human perception, in which also sites, site sections and content hierarchies can be distinguished. This abstraction is inherent to current search engines as Google, Bing or Yandex, which are optimized to return the most interesting individual documents as a ranked list (they do not return the most interesting sites or site sections). Thus, the largest part of research on search engines addresses the fundamental problem of composing the result list. This includes ranking (e.g. [5]), diversification (e.g. [6]) and personalization (e.g. [7]). Other work focusses on the presentation of search results, e.g. by clustering the results [8] or improving the visual arrangement [9], but to our knowledge, integrating taxonomy information as proposed in this paper has not been discussed before. However, the search engine Google does already integrate taxonomy information in another way, by displaying mined breadcrumb trails (cf. Figure 1). Details about the mining method are not published. In contrast, so-called deep-links as supported by Google and Bing (cf. Figure 4 (1)), are not based on the original Web site taxonomy, but on ranking algorithms.

A lot of work exists on extracting or generating taxonomies from Web content, usually based on text analysis, hyperlink structure, URL structure or a combination of these features (e.g.[10],[11],[12]). However, these works do not aim at recovering the original taxonomies as designed by the information architects (cf. definition in Section 3). In our previous work we were presenting novel methods to close this gap [3].

Semantic search is a research topic that aims at utilizing the Web of Data for augmenting traditional search with additional information, determining the context of a query based on domain knowledge or overcoming the traditional document-centric approach by query answering technologies. As Guha et al. [13] have observed, there are two kinds of searches, navigational searches and research searches. They argue that semantic search attempts to improve research searches. Since the Web of Data is a promising source of taxonomy information, our paper shows that rich markup can be used to improve navigational searches, too.



Transformation - W3C

www.w3.org › Standards › XML Technology

Site Navigation. **XML Technology** Header link ... XSLT and XSL-FO a

Recommendations for defining **XML** document transformation and pres

Figure 1. Search result with breadcrumb trail (google.com)

Using click-through data to improve ranking is a common approach, but few works consider data from user interaction with the linked resources. A detailed analysis of different features capturing the searcher's behavior on the landing page can be found in [14]. Complete post-click navigation trails have been studied and it was shown that users do not only benefit from the information on the landing page and the destination, but also from the intermediate pages [15][16]. To our knowledge, the influence of taxonomies on post-search navigation was not studied before.

3. WEB SITE TAXONOMIES

In Web-related research, the terms hierarchy and taxonomy are sometimes used as synonyms, sometimes used with different meanings. Based on an information architecture point of view, we propose the following definition, which is used in the rest of this paper:

Definition: *The term Web site taxonomy denotes labels for a group of Web resources of the same site and a logical tree structure in which they are arranged. The labels and the tree structure are designed with the purpose of facilitating access to resources. Each label describes the associated document and the documents associated with all descendant labels.*

Thus, in this paper Web site taxonomies are understood as the logical organization behind the hierarchical menus that can be found on almost all Web sites. The nodes of the tree structure are given by labels, but each label is associated with a document. Labels that are not leaves represent multiple resources. For example, the node "sports" in a Web site taxonomy of a news site subsumes a large number of individual documents associated with descendant nodes, e.g. "soccer". Thus, the tree edges can be interpreted as type-subtype relationships and we prefer the term taxonomy to hierarchy. Web site taxonomies as understood in this paper are logical structures, not link structures.

To understand the role taxonomies play for accessing Web sites, it is helpful to switch to the perspective of usability experts. There is a broad agreement on the importance of taxonomies (i.e., hierarchies) for Web site organization. Well-designed taxonomies are the "foundation of almost all good information architectures" [1], hierarchical structures are "far and away the most common" [2] so that "most Web sites have some kind of hierarchy" [17]. The whole chapter about Web navigation in [18] assumes an underlying taxonomy as a matter of course. As noted in [1] the preference of hierarchies may seem "blasphemous" in an hypertextual environment, but makes sense from the designer's or information architect's perspective. Taxonomies are familiar and humans "have been organizing information into hierarchies since the beginning of time" [1]. The authors of [17] and [18] mention two concepts to indicate the user's position: These are highlighting the active menu item and using breadcrumb trails. Both concepts are based on an underlying hierarchical content organization. From the Web designer's perspective the idea of locality seems to be naturally connected to taxonomies. For example, a user cannot answer where a certain article is located on the Wikipedia site because it is one of the few sites that are not relying on a hierarchical organization. But hierarchies allow users to develop a mental model for the site organization and their location [1].

4. ENHANCING SEARCH RESULT SNIPPETS WITH TAXONOMY DATA

Analysis of logged search trails shows that after a click on a promising search result a phase of navigation on the target site often follows (e.g. [15],[16]). Teevan et al. state that "the perfect search engine is not enough" and post-search navigation is not due to limitations of the search engine but due to human search behavior [19]. The authors of the study found that searchers often prefer to enter less specific keywords (e.g. the domain name of the target site if known) to narrow in on the target page afterwards by a series of navigation steps, which results in a lower cognitive load compared to entering a more specific search query. This behavior is called orienteering by the authors. Another explanation for the frequency of post-search navigation besides the orienteering theory is that the information need is often more complex and cannot be satisfied by a single page alone [15].

The post-search navigation scenario is illustrated in Figure 2. (1.) By entering the search query the user is transferred to the Search Engine Result Page (SERP). (2.) Then the user clicks on a promising result and is transferred to the landing page. (3.) The user assesses the information found on the landing page and decides if either to end the search, to return to the SERP or to continue browsing on the target site. Current search engines display short summaries of the target pages on the SERP, called search result snippets. By this they are trying to give hints whether the user can satisfy his information need on the target page.

We argue that with post-search navigation in mind the focus of the search result snippets should be extended. Instead of summarizing the content of the target pages only, search result snippets should also try to answer the question whether a certain target page is a good starting point for further exploration. At the same time *"Where am I?"*, *"What's here?"* and *"Where can I go next"* are the three basic questions Web navigation can answer [17]. The question *"Where can I go next"* is equivalent to *"Is this page a good starting point for further exploration?"*. Because of that, information about the navigation options on the target page can be very useful in search result snippets – and as we have argued in Section 3, taxonomies play a central role for Web navigation. Thus, taxonomy information can be especially valuable for search result snippets.

To illustrate this idea we created two mockups. Figure 3 shows the summary for the page "Participants" from the official WWW2013 Web site as it is displayed by google.com. The text snippet originates from the provided HTML metadata. In the mockup below we replaced the text summary by links to the child nodes of the target page. In contrast to the original presentation the child nodes provide very useful information about the navigation target. We do not propose to replace all text snippets with taxonomy information – the example is to illustrate the usefulness of child node information. Child nodes and text snippets can be used in combination.

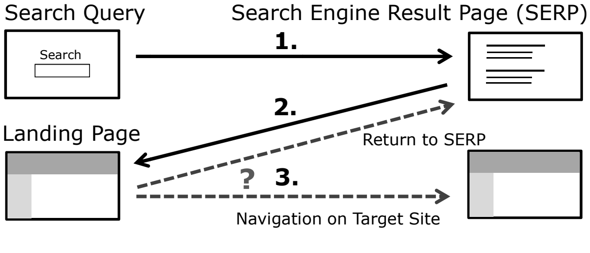
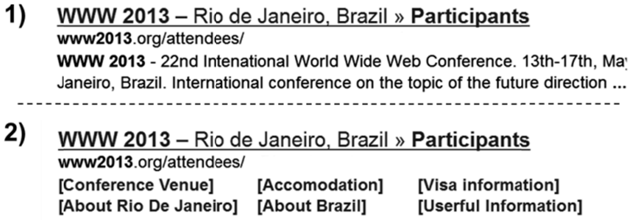


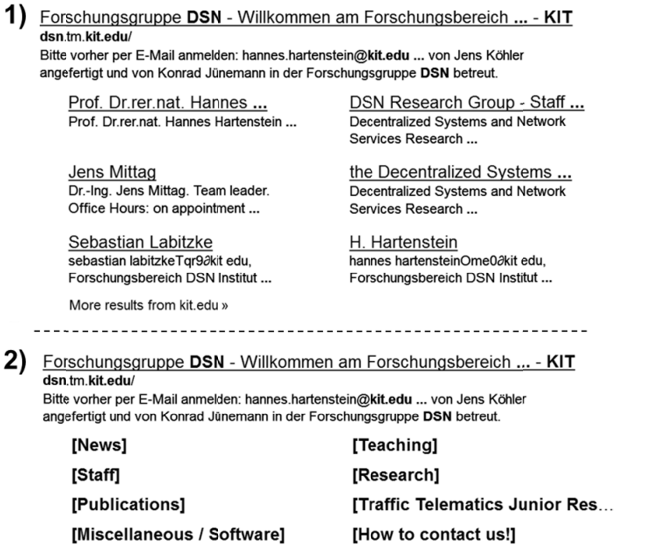
Figure 2. Scenario: After the search results are displayed (1), the user is transferred to the landing page of a target site by clicking on a promising result (2). The user ends his search, returns to the SERP or continues browsing (3).



1) WWW 2013 – Rio de Janeiro, Brazil » **Participants**
www2013.org/attendees/
WWW 2013 - 22nd Intenational World Wide Web Conference. 13th-17th, Ma
Janeiro, Brazil. International conference on the topic of the future direction ...

2) WWW 2013 – Rio de Janeiro, Brazil » **Participants**
www2013.org/attendees/
[Conference Venue] [Accomodation] [Visa information]
[About Rio De Janeiro] [About Brazil] [Userful Information]

Figure 3. (1) Search result snippet for the page "Participants" from the official WWW2013 Web site as presented by google.com[2]. The displayed text is the description provided by the corresponding meta-tag. (2) Mockup in which the description is replaced by links to the child pages. While the original presentation does provide few useful hints about the navigation target, the mockup gives a very good summary.



1) Forschungsgruppe **DSN** - Willkommen am Forschungsbereich ... - **KIT**
dsn.tm.**kit.edu**/
Bitte vorher per E-Mail anmelden: hannes.hartenstein@**kit.edu** ... von Jens Köhler
angefertigt und von Konrad Jünemann in der Forschungsgruppe **DSN** betreut.

Prof. Dr.rer.nat. Hannes ...
Prof. Dr.rer.nat. Hannes Hartenstein ...

DSN Research Group - Staff ...
Decentralized Systems and Network Services Research ...

Jens Mittag
Dr.-Ing. Jens Mittag. Team leader. Office Hours: on appointment ...

the Decentralized Systems ...
Decentralized Systems and Network Services Research ...

Sebastian Labitzke
sebastian labitzkeTqr9∂kit edu, Forschungsbereich DSN Institut ...

H. Hartenstein
hannes hartensteinOme0∂kit edu, Forschungsbereich DSN Institut ...

More results from kit.edu »

2) Forschungsgruppe **DSN** - Willkommen am Forschungsbereich ... - **KIT**
dsn.tm.**kit.edu**/
Bitte vorher per E-Mail anmelden: hannes.hartenstein@**kit.edu** ... von Jens Köhler
angefertigt und von Konrad Jünemann in der Forschungsgruppe **DSN** betreut.

[News] [Teaching]
[Staff] [Research]
[Publications] [Traffic Telematics Junior Res...
[Miscellaneous / Software] [How to contact us!]

Figure 4. (1) Google supports shortcut links to pages that are considered as the most important pages of a site.[3] The linked pages are computed using Google's ranking algorithms and do not reflect the main topics of the site. (2) The mockup below shows an alternative presentation for the site based on the taxonomy. Users are able to get an overview of the content, even if the text snippet is not understandable.

Current search engines already enrich search snippets with additional links, but they use different approaches. The so called *deep-links* or *site-links* are links to individual pages that are considered as the most important pages based on the ranking algorithm and of the search engine and click logs. They do not reflect the logical structure of a site and they do not contain information about the sub-pages of a linked search result. Figure 4 (1) shows the search result snippet displayed by Google representing our research group's Web site. Links to the English

[2] Source page: http://www2013.org/attendees, retrieved on 2013/2/14

[3] Source site: http://dsn.tm.kit.edu/, retrieved on 2013/2/14

pages of individual group members and alumni students are displayed as deep-links, while the German entry page is presented above. The deep-links might represent the most-visited pages of the site but they do not give information about the complete range of information that can be found there, in contrast to the mockup with child-nodes below. The deep-links are helpful for users that are seeking for information that can be found on one of the six linked pages, but the child-links are useful for all users interested in the target site. The example illustrates the idea of considering post-search navigation in Web search. While current search engines try to teleport users to certain documents, the taxonomy approach includes more information about what a user can find on a site or a site section.

5. EVALUATION

The idea of integrating taxonomy information in the presentation of search results is based on the findings that post-search navigation is common and, moreover, usability experts state that taxonomies play a central role in Web navigation. However, the influence of taxonomies on post-search navigation has not been empirically researched yet. In this section we present a study that answers the following two questions:

- Can we observe a preference for child-links compared to other links in post-search navigation (Figure 5)?
- Are there situations in which child-links are more effective as shortcut links than deep-links provided by current search engines?

The two questions result from the fact that there are two ways users can benefit from the integration of child nodes: The first scenario is that child nodes help users to estimate whether it is worth visiting a linked search result, because they provide additional information about the navigation target. The first question aims at verifying that the child-nodes of the landing page are the preferred next navigation steps, and, thus, provide useful information. The second scenario is that child nodes are used as shortcuts to directly access them without visiting the parent. The second question aims at analyzing, whether child nodes can be effective shortcut links in comparison with deep-links used by current search engines.

5.1 Experimental Setup

To gain insights in the role of taxonomies in post-search navigation, clickstream-data from a search engine is not sufficient. In addition, the interaction with the target sites has to be tracked. Such kind of data is available to providers of browser toolbars and site owners. Moreover, the taxonomies of the target sites must be extracted and the necessary markup annotations or mining methods are not yet broadly available. Thus, we analyzed the usage data of Web sites of which we could access the server logs and extract the Web site taxonomy from the underlying content management database.

We extracted clickstreams from the server log files of three Web sites for a period of two weeks in May 2012. We considered the sites www.kit.edu (A), www.scc.kit.edu (B) and the site of a municipality responsible for about 25,000 citizens (C). The clickstreams were preprocessed to remove entries from crawlers and to identify individual sessions. At the end 470,827 clickstreams for Site A were analyzed, 89,360 for Site B and 15,953 for Site C. The clickstreams were aligned with a model of the content hierarchy to analyze the navigation behavior in relation to it. At that time the hierarchies consisted of 628 (A), 632 (B) and 154 (C) elements. They had a maximum depth of 10 (A), 9 (B) and 4 (C).

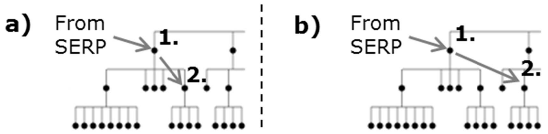


Figure 5: a) User visiting a child node after the landing page and b) user visiting a linked non-child node

5.2 Preference of Child-Links

For a Web site $W=\{w_1, w_2 \dots w_n\}$ consisting of n pages w_1, w_2... w_n, the taxonomy is defined by sets $C_{w_i} \subseteq W$ of child pages associated with a parent page w_i. A set of j clickstreams $S = \{S_1, S_2 \dots S_j\}$ is given. The k-th clickstream $S_k = \{s_{k,1}, s_{k,2}, \dots s_{k,z_k}\}$ describes a user session consisting of z_k consecutive clicks, where $s_{k,l} \in W$ for $l=1 \dots z_k$. Regarding the post-search navigation scenario, we are interested in users that arrive from a search engine (which can be identified by the HTTP referer) on a landing page and continue browsing on the site. If $1_A(x)$ is the indicator function that has the value 1 if $x \in A$ and 0 otherwise, the ratio of clickstreams RC_S whose second visited page is a child node of the first visited page is given by:

$$RC_S = \frac{1}{|S|} \sum_{m=1}^{|S|} 1_{C_{s_{m,1}}}(s_{m,2})$$

Still that ratio would be heavily biased by the ratio of child node links on the landing page. Thus, we normalized ratio AC_S by dividing by the average number of child nodes on the landing pages:

$$AC_S = \frac{\frac{1}{|S|} \sum_{m=1}^{|S|} 1_{C_{s_{m,1}}}(s_{m,2})}{\frac{1}{|S|} \sum_{m=1}^{|S|} |C_{s_{m,1}}|} = \frac{\sum_{m=1}^{|S|} 1_{C_{s_{m,1}}}(s_{m,2})}{\sum_{m=1}^{|S|} |C_{s_{m,1}}|}$$

In other words AC_S is the ratio of the number of all clicks on child nodes of the landing pages to the number of all child nodes that were shown on the visited landing pages. The same way we computed AC_S^- as the ratio of clicks on non-children to the number of all shown non-child links. By comparing both values we can estimate if it is more likely that a user clicks on a certain link that leads to a child page of the landing page than clicking on a certain link which does not. For both AC_S and AC_S^- we only considered clickstreams with length > 1 and landing pages that belong to the Web site taxonomy and had child nodes.

Table 1 shows that in average child-links are much more likely to be clicked from users arriving from a search engine than non-child links. The ratio of child link clicks to non-child-link clicks ranges from about 4.7 to 11.7. Because of the preference for child links it can be concluded that users clearly tend to descend in the content hierarchy from the landing page. The results are very much in accordance with the idea of post-search orienteering and narrowing in on the target [19] – which is equivalent to moving down the Web site taxonomy towards more specific results. To see whether the behavior of users on search engine landing pages differs from the general browsing behavior, we also included clicks that did not follow a search. The extended data set shows a similar preference for child links.

Regarding the search engine scenario, it can be concluded that displaying the child nodes on the SERP would anticipate the options for the next browsing step effectively, at least for similar types of sites. The findings indicate that the extra information would help users to pre-estimate whether a search result is a good starting point for further exploration and worth visiting.

Table 1: Child node hits vs. non-child node hits

	Hits from Search Engines				All Hits			
	#Clicks	AC_s	$AC_{\bar{S}}$	$\frac{AC_s}{AC_{\bar{S}}}$	#Clicks	AC_s	$AC_{\bar{S}}$	$\frac{AC_s}{AC_{\bar{S}}}$
Site A	14918	0.0681	0.0058	11.74	91520	0.081	0.009	9
Site B	1672	0.0672	0.0143	4.7	19017	0.074	0.013	5.69
Site C	736	0.0925	0.0085	10.88	3525	0.068	0.012	5.67

Table 2: Parent node hits vs. non-parent node hits

	Hits from Search Engines				All Hits			
	#Clicks	AP_s	$AP_{\bar{S}}$	$\frac{AP_s}{AP_{\bar{S}}}$	#Clicks	AP_s	$AP_{\bar{S}}$	$\frac{AP_s}{AP_{\bar{S}}}$
Site A	3711	0.0182	0.0189	0.96	93858	0.019	0.024	0.79
Site B	2031	0.0429	0.0259	1.66	27930	0.04	0.021	1.9
Site C	294	0.068	0.021	3.24	3908	0.044	0.023	1.91

The same metrics were computed to analyze the preference for parent nodes (Table 2). Interestingly only on two out of the three sites, links to parent pages are clicked more often than random links. The preference for child nodes is far more significant.

5.3 Child-Links as Shortcut-Links

If users are not interested in the information on the landing page itself, child nodes can also help saving clicks by providing shortcuts to more specific results. To get an idea of the potential of child nodes as shortcuts we compared child nodes to shortcut links provided by current search engines. Those are called “deep-links” or “site-links” and are representing the most relevant pages of a site as ranked by the search engine, usually based on link analysis algorithms as PageRank [4] and click-through rates. Deep-links are presented in combination with the entry page of a site (cf. Figure 4).

From now on it is assumed that a search engine has a perfect ranking algorithm that delivers an accurate list of the most interesting pages of a site. Furthermore, we assume that the search engine provides the top entries of the list as additional shortcut links if a page of the site is returned as result. We model the list of the most interesting pages based on our usage data. For this we generated a list of pages ordered by the number of hits they received for each site from the log files. The clickstreams were then processed to compute the ratio of users that proceeded to a first child node of the landing page to compare it to the ratio of users that proceeded to the page with the site-wide most visits (Figure 6). This was done for all clickstreams with length > 1, if the landing pages had child nodes. For the same clickstreams, the ratio of clicks that were received by the linked page with the site-wide most hits was computed. In a similar way all clickstreams where the landing page had at least 2 child nodes were processed, to compare the ratio of clicks on the first two children to the ratio of clicks that the two linked pages with the best global ranking received. This was done for all three sites up to a child number of six (Figure 7).

The plot can be interpreted as follows: If, e.g., four shortcut links per search result shall be presented, the deep-links based on a global ranking would be the better choice for Site A, because in average about 54 percent of the first clicks on the landing page were attracted by those four pages, while the four first child nodes received in average about 41 percent of the clicks. However, for Site B the first four child nodes of the landing pages received in average about 51 percent of the clicks versus only 35 percent that were attracted by the globally ranked links. For Site C again, the globally ranked nodes attract slightly more clicks.

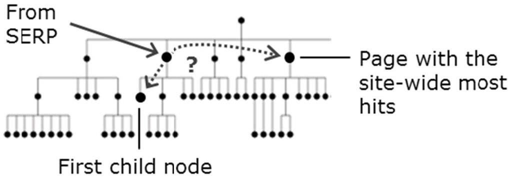


Figure 6: How does the number of users that proceed to the first child node of a landing page compare to the number of users that proceed to the page with the site-wide most visits?

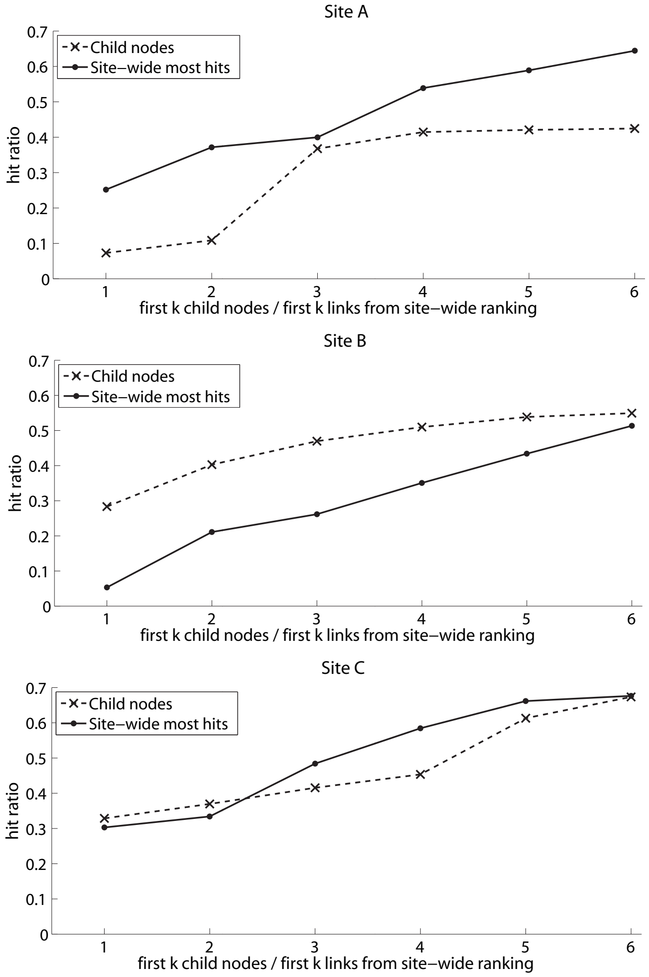


Figure 7. Child links vs. site-wide ranking: For each site the ratio of hits received by the first k child nodes and the ratio of hits received by the k linked pages with the site-wide most hits are plotted.

The results indicate that links to child nodes of search results can be very useful as shortcuts, sometimes even more attractive than deep-links. However there is no need to trade-off deep-links and links to child pages. Both complement each other. Deep-links can be used to provide shortcuts to the most important contents for whole sites while child page links provide shortcuts to the next most interesting browsing destinations for individual pages.

5.4 Summary

Based on the results of the experiment we can conclude:

- Taxonomies have a strong influence on where users go in post-search navigation. Empirical findings show that users arriving from a search engine are much more likely to navigate to a child page than to another random linked page.

We conclude that child links would help users to estimate if post-search navigation promises useful information.

- The findings indicate that child nodes can provide useful shortcuts that can be generated without statistical click data. They can be more relevant for a user's search than the shortcut links provided by current search engines and can be a better model for the next navigation steps.
- The experiments show that even in case of informational and hierarchically organized sites the navigation behavior in relation to the taxonomy differs. More research is necessary to clarify for which sites child links are suitable and how they can be combined with other features.

6. CONCLUSION AND FUTURE WORK

In this paper we have demonstrated how Web site taxonomies, which are increasingly available to machines, can be utilized to further improve the presentation of search results. We have analyzed that not only the target document should be summarized in search result snippets but also the most interesting options for the next navigation steps, which are the child pages in case of taxonomies. An empirical study supports our argumentation. By analyzing three highly-frequented Web sites we have shown that a link to a child node of a landing page receives up to 11 times as much hits as a random other link in post-search navigation. Thus, child nodes allow users to anticipate the next navigation options effectively and help them assessing whether a landing page is a good starting point for further exploration. We have also shown that child nodes are a promising complement to shortcut links current search engines provide based on their ranking algorithms.

An analysis of the question whether the findings generalize to all sites with taxonomies or just some sites with taxonomies was out of scope of this paper. Strategies for deciding in which cases taxonomy information should be presented, exclusively or in combination with other features, are interesting for future research, as well as the question, which metrics are suitable.

Another possible extension of the presented ideas is that, if the ranked list of search results contains multiple documents with the same parent, the search results can be presented aggregated and more clearly arranged. Instead of displaying all child nodes separately the parent node with child links could be presented, even if the parent itself does not match the search query.

Search engines could start integrating taxonomy information of sites that have semantically annotated breadcrumb trails, according to the vocabulary of schema.org. From the breadcumb trails of individual pages the complete site taxonomy can easily be assembled. If search engines would start utilizing taxonomy information, Web developers in turn would be motivated to integrate the corresponding semantic markup in their code.

7. REFERENCES

[1] P. Morville and L. Rosenfeld, *Information architecture for the World Wide Web*. Sebastopol, (Calif.) [etc.]: O'Reilly, 2006.

[2] J. J. Garrett, *The elements of user experience : user-centered design for the Web and beyond*. Berkeley, CA: New Riders, 2011.

[3] M. Keller and M. Nussbaumer, "MenuMiner: revealing the information architecture of large web sites by analyzing maximal cliques," in *Proceedings of the 21st international conference companion on World Wide Web*, New York, NY, USA, 2012, pp. 1025–1034.

[4] L. Page, S. Brin, R. Motwani, and T. Winograd, *The PageRank Citation Ranking: Bringing Order to the Web*. 1999.

[5] C. Kang, X. Wang, Y. Chang, and B. Tseng, "Learning to rank with multi-aspect relevance for vertical search," in *Proceedings of the fifth ACM international conference on Web search and data mining*, New York, NY, USA, 2012, pp. 453–462.

[6] D. Rafiei, K. Bharat, and A. Shukla, "Diversifying web search results," in *Proceedings of the 19th international conference on World wide web*, New York, NY, USA, 2010, pp. 781–790.

[7] D. Sontag, K. Collins-Thompson, P. N. Bennett, R. W. White, S. Dumais, and B. Billerbeck, "Probabilistic models for personalizing web search," in *Proceedings of the fifth ACM international conference on Web search and data mining*, New York, NY, USA, 2012, pp. 433–442.

[8] U. Scaiella, P. Ferragina, A. Marino, and M. Ciaramita, "Topical clustering of search results," in *Proceedings of the fifth ACM international conference on Web search and data mining*, New York, NY, USA, 2012, pp. 223–232.

[9] F. Chierichetti, R. Kumar, and P. Raghavan, "Optimizing two-dimensional search results presentation," in *Proceedings of the fourth ACM international conference on Web search and data mining*, New York, NY, USA, 2011, pp. 257–266.

[10] R. Kumar, K. Punera, and A. Tomkins, "Hierarchical topic segmentation of websites," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, 2006, pp. 257–266.

[11] W. K. Cheung and Y. Sun, "Identifying a hierarchy of bipartite subgraphs for web site abstraction," *Web Intelli. and Agent Sys.*, vol. 5, no. 3, pp. 343–355, Aug. 2007.

[12] Q. Ho, J. Eisenstein, and E. P. Xing, "Document hierarchies from text and links," in *Proceedings of the 21st international conference on World Wide Web*, New York, NY, USA, 2012, pp. 739–748.

[13] R. Guha, R. McCool, and E. Miller, "Semantic search," in *Proceedings of the 12th international conference on World Wide Web*, New York, NY, USA, 2003, pp. 700–709.

[14] Q. Guo and E. Agichtein, "Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior," in *Proceedings of the 21st international conference on World Wide Web*, New York, NY, USA, 2012, pp. 569–578.

[15] R. W. White and J. Huang, "Assessing the scenic route: measuring the value of search trails in web logs," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 2010, pp. 587–594.

[16] A. Singla, R. White, and J. Huang, "Studying trailfinding algorithms for enhanced web search," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 2010, pp. 443–450.

[17] J. Kalbach, *Designing Web navigation*. Beijing; Sebastopol: O'Reilly, 2007.

[18] S. Krug, *Don't make me think! : a common sense approach to Web usability*. Berkeley, Calif: New Riders Pub., 2006.

[19] J. Teevan, C. Alvarado, M. S. Ackerman, and D. R. Karger, "The perfect search engine is not enough: a study of orienteering behavior in directed search," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2004, pp. 415–422.