Mining User-generated Path Traversal Patterns in an Information Network

Frank W. Takes and Walter A. Kosters
Leiden Institute of Advanced Computer Science (LIACS), Leiden University
{ftakes,kosters}@liacs.nl

***Abstract*—This paper studies patterns occurring in user-generated clickpaths within the online encyclopedia Wikipedia. The clickpath data originates from over seven million goal-oriented clicks gathered from the Wiki Game, an online game in which the goal is to find a path between two given random Wikipedia articles. First we propose to use node-based path traversal patterns to derive a new measure of node centrality, arguing that a node is central if it proves useful in navigating through the network. A comparison with centrality measures from literature is provided, showing that users generally "know" only a relatively small portion of the network, which they employ frequently in finding their goal, and that this set of nodes differs significantly from the set of central nodes according to various centrality measures. Next, using the notion of subgraph centrality, we show that users are able to identify a small yet efficient portion of the graph that is useful for successfully completing their navigation goals.**

***Keywords*-path traversal; navigation; centrality; information networks; Wikipedia**

I. Introduction

A large part of the gigantic amount of information that is nowadays available is organized in some sort of *network* structure. Examples include the world wide web, an online social network or an information network such as Wikipedia. In these networks (or graphs), each node represents an entity or a piece of information, and each link represents a tie or relationship between two entities. An important task that human users perform on a daily basis, is *searching* for a piece of content within such a network. Although search engines can often assist the user in performing such a search task, *navigating* to the desired page by means of clicking the links between the nodes in the network is still a common activity, as sometimes search engine performance does not exactly meet the user's needs [1]. This can happen because the user's query is misinterpreted, or because the required information is not indexed, for example because it is located within the so-called "Deep Web" [2]. In such cases, the user will have to reach the correct page by traversing hyperlinks that exist between the pages in the network, forming a path towards the correct piece of information. Throughout this paper we consider the task of mining *traversal patterns* that occur within these types of clickpaths in an attempt to better understand human search behavior.

The path traversal data used in this research originates from the Wiki Game, an online game in which the main task is to link two given random Wikipedia articles. Employing his perception of the structure of the network, a user has to find his way to the goal article by clicking the directed links that exist between the various articles in the Wikipedia graph, essentially generating a goal-oriented clickpath. We will consider more than one million of these clickpaths, containing over seven million clicks on Wikipedia pages. It is important to note that these clicks are fundamentally different from simply counting the number of visits to a certain page, as these counts would for example also include visits that immediately reach the desired goal page, for example via a search engine. Instead, the clickpaths that we will study consist of Wikipedia pages and links between pages that were actually considered useful, by the user, in *traversing* the Wikipedia network.

We will use node-based traversal patterns to address a problem within the field of network analysis called *node centrality*, defined as the importance of a node within the network. So-called *centrality measures* are widely used to assess this issue of node centrality. One of the most well-known examples of such measures is PageRank [3], which assigns a score between 0 and 10 to a webpage, indicating the importance of this webpage with respect to the rest of the web. Other commonly used centrality measures originate from the field of social network analysis, and include degree centrality, closeness centrality and betweenness centrality [4]. While the aforementioned centrality measures all employ the structure of the network to assess the importance of a node, none of them incorporates the human perception of the information incorporated in the network. As it is ultimately the user who is going to assess whether or not a page is actually relevant, one could say that it is not the structure of the network which should serve as the basis of the centrality measure, but it should instead be the user's perception of the network that is going to determine the importance of a node. It may very well be that certain structurally central nodes in the network are not at all considered important or useful by the user, and vice versa. Therefore we introduce a user-defined measure of centrality based on frequently traversed nodes, arguing that a page is important if it proves useful in navigating through the network. Especially in networks where the user perception of the data plays a central role, such as in the world wide web, or in an information network, we believe that a user-defined measure makes more sense than a conventional user-insensitive approach. Furthermore, we introduce the measure

of subgraph centrality which determines the centrality of a group of connected nodes with respect to the rest of the network, allowing us to experimentally verify the quality in terms of ease of navigation of the user-perceived central nodes.

The rest of the paper is organized as follows. In Section II we discuss some definitions and introduce our dataset. After discussing related work in Section III, we consider node-based patterns and our user-defined measure of centrality in Section IV. We assess its performance by means of experiments on the Wikipedia and Wiki Game dataset in Section V. Finally, Section VI concludes the paper and provides suggestions for future work.

II. Preliminaries

This section starts with some basic definitions regarding graphs and paths that will later on allow us to precisely define our path traversal patterns and various derived measures. We also describe the clickpath dataset to which we will apply our path traversal pattern mining techniques.

A. Definitions

We will model the information network Wikipedia as a directed graph $G(V, E)$ with $n = |V|$ nodes and $m = |E|$ directed links between pairs of nodes. The indegree $indeg(v)$ of a node $v \in V$ is equal to the number of incoming links of v, and similarly $outdeg(v)$ denotes the number of outgoing links. We define a *path* as a sequence $(v_1, v_2, \ldots, v_\ell)$ of ℓ visited nodes, where for each consecutive node pair there exists a link $(v_i, v_{i+1}) \in E$ (with $1 \leq i < \ell$) in graph G. The *path length* is then equal to $\ell - 1$, the number of links that were traversed to get from the first to the last node in the path. We define the *distance* $d(u, v)$ as the length of the shortest path between nodes u and v, meaning the minimum number of links that has to be traversed to get from u to v. If there is no path between u and v, then $d(u, v) = \infty$. In such cases, the graph has multiple *strongly connected components*, meaning that some nodes are not reachable from every other node by considering the directed links between the nodes. Similarly, in a *weakly connected component* there is a path from each node to every other node in the component, ignoring the direction of the links. For convenience in later definitions, we denote the number of shortest paths by $\sigma(v, w)$, and the number of shortest paths from v to w that run through node u by $\sigma_u(v, w)$.

B. Wikipedia dataset

In this research, we use a Wikipedia graph consisting of the pagelinks from the English version of DBPedia 3.7 and 3.8 [5] that was mined from the original Wikipedia website in 2011 and 2012. We mention that by only considering actual pagelinks and ignoring links to special pages or external websites, each page represents an actual piece of information within the information network. The graph consists of $n = 3,416,126$ nodes and $m = 83,271,539$ directed links, and has a large weakly connected component consisting of 99.98% of the total number of nodes. The degree distribution follows a power law, and the distance distribution and average node to node distance (4.55) are consistent with that of other small world networks [6], meaning that the structure of the Wikipedia graph is on a global scale somewhat similar to that of for example social networks or web graphs.

C. The Wiki Game dataset

The clickpath data used in this paper is based on clicks on Wikipedia articles made by users of the Wiki Game (`http://www.thewikigame.com`), an online game which was introduced in 2009. In this game, users are assigned the task of connecting two given random articles on Wikipedia by traversing the links that exist between Wikipedia articles.

The original dataset consists of clickpaths generated between 2009 and 2012, where one clickpath corresponds to a played game (or task), which is is essentially a (start, goal)-pair in between which a path has been formed. A total of $527,300$ different users have attempted to solve such a task, generating $3,219,641$ paths consisting of $17,151,824$ clicks in total. Of these tasks, little over one third was successfully completed. In this paper, we will only consider successful paths (won games), with a length between 3 and 20, thus filtering out non-serious attempts and failed clickpaths. This results in a dataset of $1,137,337$ clickpaths consisting of a total of $7,135,060$ clicks.

As an example of a path traversal task, consider the path from the Wikipedia article on Sleep Disorder to the article on Quebec (the Canadian province). Figure 1 shows a subgraph of the Wikipedia graph based on six paths generated by six different users that successfully solved this task, aggregating links that were traversed more than once by increasing the width of the link. Most users first somehow find their way to a geography-related page, after which they, often taking a detour via the page on Quebec city, traverse to the actual article on the identically named province. While one path has the actual optimal shortest path length of four, most users take detours and use quite a few more steps to find the goal page. This is also demonstrated in Figure 2, in which we observe tailed distributions of the lengths of all performed pathfinding tasks, as well as that of the computed shortest paths lengths. For more information on (an older and smaller version of) the Wiki Game dataset, we refer the reader to our previous paper [7].

III. Related work

Path traversal patterns in a hyperlinked environment have been a popular subject of study since the introduction of the web, and a lot of work has been done on mining the top-k frequent traversal patterns [8]. In a web setting, studying

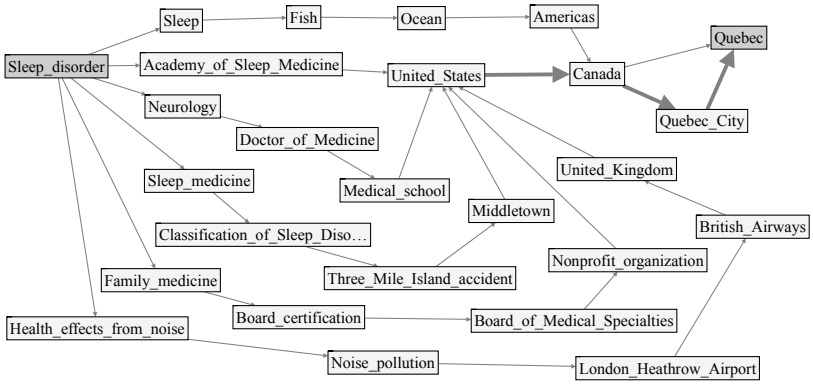

Figure 1. Subgraph of Wikipedia based on six user-generated paths from the article on Sleep Disorder to the article on Quebec.

path traversal patterns from a stream is also a relevant task [9]. Most research in which clickpaths are analyzed within a confined environment deals with pages from a particular website [10]. We distinguish from such studies, because first of all, every click in our dataset is goal-oriented and second, clicks are always identifiable as one unique topic, namely the subject of the Wikipedia page.

West and Leskovec [11] have compared human navigation in information networks such as Wikipedia with that of agents, using a dataset similar to ours. They found that humans, when navigating within an information network, have expectations about what links should exist and base a high level reasoning plan upon this, and then use local information to navigate through the network. They furthermore mention that humans often miss "good" link opportunities on a page as their idea of semantic relatedness often overrules opportunistic clicking. In [12], the same authors show that progress in a goal-finding task is easiest far from and close to the target, with hubs being crucial in the beginning.

In [7], we have investigated the difficulty of forming a path between two given random pages, showing that in the Wiki Game, the indegree of the goal page as well as the reversed neighborhood, both local properties of the goal page, are good predictors of the difficulty of performing such a path traversal task. We have also demonstrated how the start page is of little influence as the user just navigates away from it quickly in search for a hub. Whereas the previous paper only considered path success or failure, in this paper, we consider the patterns that arise from the actual clicks made by the users.

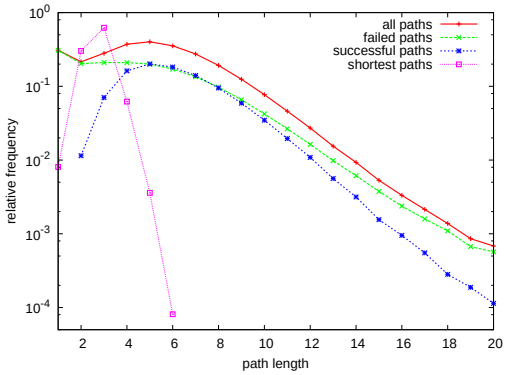

Figure 2. Relative frequency (vertical axis, logarithmic) of various path lengths (horizontal axis) over all user-generated paths in the Wiki Game.

IV. Path traversal patterns

In this section we will first introduce three types of path traversal patterns, after which we look in detail at node-based traversal patterns, and how these patterns can serve as a basis of a user-defined measure of centrality. In the next section, we will compare this new measure with centrality measures from literature.

A. Patterns

Given a set P consisting of a large number of clickpaths, we are interested in *patterns*, i.e., observable phenomena that occur more frequently than normal. Similar to the definitions often given in the area of frequent itemset mining, we call an observation *frequent* if it occurs more often than a certain threshold $\theta > 0$ amongst all paths. We then define the set of top-k frequent patterns as the set of $k \geq 1$ patterns with the highest frequency. For our clickpath dataset, we will distinguish between the following patterns:

- *Top-k frequent nodes*: the k most frequently visited nodes in all paths $p \in P$.
- *Top-k frequent edges*: the k most frequently traversed pairs of consecutive nodes in all paths $p \in P$.
- *Top-k frequent subpaths*: the k most frequently traversed ordered sequences of three or more consecutive nodes in all paths $p \in P$.

Obviously, relaxing the definition of frequent subpaths to length two or one, yields the definitions of respectively edge and node traversal frequency. As an example, Figure 3 shows the frequency of each node and edge traversal count over all nodes in the graph. The most simple patterns based on frequent nodes are further discussed in Section IV-C.

B. Centrality measures

The measure of node centrality is defined as the importance of a certain node in the graph. A centrality measure M

returns the centrality $C_M(v)$ of a node v. We will consider the following five (existing) centrality measures:

$$\text{Indegree centrality} \quad C_{indeg}(v) = \frac{indeg(v)}{n-1}$$

$$\text{Closeness centrality} \quad C_c(v) = \frac{1}{\frac{1}{n-1}\sum_{w \in V} d(v,w)}$$

$$\text{Betweenness centrality} \quad C_{bc}(v) = \sum_{\substack{u,w \in V \\ u \neq w, u \neq v, v \neq w}} \frac{\sigma_v(u,w)}{\sigma(u,w)}$$

We also consider Pagerank $C_{pr}(v)$ and HITS $C_{hits}(v)$. In the PageRank measure [3], the value of $C_{pr}(v)$ is equal to $PR(v)$ after iteratively (usually 100 iterations is enough for convergence) applying the following computation for each of the nodes:

$$PR(v) = \frac{1-d}{n} + d\left(\sum_{w \in N'_v} \frac{PR(w)}{outdeg(w)}\right)$$

Here, N'_v is the set of nodes that link to node v, and d is a dampening factor, usually set to 0.15. Upon initialization, for all nodes v, $PR(v)$ is set to $1/n$.

Hyperlink Induced Topic Search (HITS) [13] is a similar technique which assigns a hub score $h(v)$ and an authority score $a(v)$ to every node v in the graph. Then, for a certain number of iterations (again 100 iterations is usually enough for convergence), each node's value of $a(v)$ is set to the sum of the (normalized) $h(u)$ values of the nodes u for which there exists a link (u,v), after which each node's value of $h(v)$ is set to the sum of the (normalized) $a(w)$ values of the nodes w for which there exists a link (v,w). For our measure of centrality C_{hits}, we use the authority score $a(v)$.

Each of the centrality measures results in a number between 0 and 1, where a higher score indicates that the node is more central. For convenience, we normalize the centrality values such that the most central node has a centrality value of 1. Clearly, distance based measures do not perform well when there is more than one connected component. Therefore we will only consider the largest strongly connected component of the Wikipedia graph when computing these measures. While more centrality measures have been developed over the past years, we believe that we have covered the most common and applicable ones in this subsection.

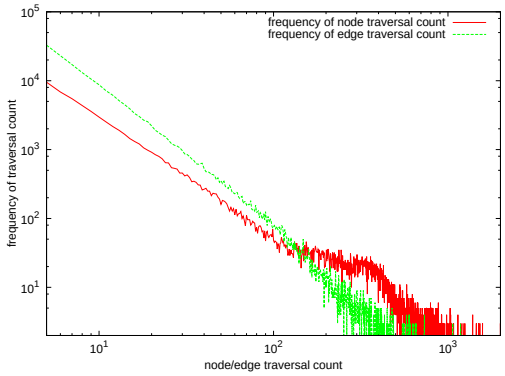

Figure 3. The frequency (vertical axis, logarithmic) of different node and edge traversal counts (horizontal axis, logarithmic).

C. User-defined node centrality

Recall from Section IV-A that considering the *top-k frequent nodes* means that if we sort the list of nodes by their node frequency value, we consider the k nodes with the highest frequency. For our clickpath dataset, this means that we are looking at the k nodes that were most frequently used to traverse the graph. This list is actually quite interesting, as it essentially indicates which k nodes are considered important, by the user, in navigating through the graph. We use this data as a basis for our user-defined measure of centrality, proposing to count the number of clicks that an article v received (denoted by *clicks(v)*) and divide it by the total number of clicks made in order to obtain our user-defined measure of centrality:

$$\text{User-defined centrality} \quad C_{ud}(v) = \frac{clicks(v)}{\sum_{w \in V} clicks(w)}$$

To get an idea of the values returned by this function, the solid line in Figure 3 shows the frequency of each node traversal count over all nodes in the graph. The distribution follows a clear power-law, meaning that many nodes are visited only a few times, and a few nodes are visited quite often. We are obviously interested in the tail of the distribution: the set of nodes that is visited very frequently.

D. Measure evaluation

Assessing the quality of a centrality measure is not a trivial task, and a manual inspection (often subjective or at least domain-dependent) is not preferred. Instead, for our experiments, we will use two automated ways of comparing centrality measures, as suggested in [14] (though in a somewhat different setting). The first rather basic technique is to compare top-k nodes of two centrality measures and determine the percentage of nodes that overlap. For example, for $k = 1$, we simply verify whether the most central node is equal for both measures. We call this measure *top-k precision*, defined as follows:

$$\text{top-}k \text{ precision} = \frac{|A_k \cap B_k|}{k}$$

Here, $A_k, B_k \subseteq V$ represent the sets of top-k nodes returned by centrality measures A and B. Second, when the actual centrality value of the top-k nodes is also of importance, we propose to look at the correlation between the centrality values in two lists of nodes. We call this evaluation measure *top-k correlation* and simply define it

as the Pearson correlation coefficient between the centrality values of the two methods.

Important to note here is that measure A is considered to be the ground truth: we compare the centrality values of the top-k nodes of measure A with the values of these nodes as determined by measure B. Finally, note that a centrality measure is used to find the top-k most central nodes, and the evaluation techniques that we discuss here are thus designed such that only the top-k nodes are evaluated.

V. Experiments

In this section we first compare the user-defined measure of node centrality introduced in Section IV-C with the existing centrality measures listed in Section IV-B. Next, we assess the quality of the various sets of central nodes using the notion of subgraph centrality in Section V-B.

A. Results

In Figure 4 we compare the different measures up to $k = 250$ using the top-k precision measure. We note that for small values of k, big deviations for the top-k precision measure can be observed, which is due to the fact that with a low value of k, one mismatch has a relatively high influence on the actual percentage. In our experiments we found that it is important not to ignore the directed aspect of the Wikipedia network, as otherwise overview pages containing listings of events or people will be ranked too high. This is also the reason why both outdegree centrality and the HITS algorithm using the hub score instead of the authority score did not produce meaningful results. Table I shows the top-k correlation, allowing us to conclude that PageRank gives not only the highest, but judging from Figure 4 also gives the most consistent results when top-k precision is considered. Indegree centrality is a good second choice if top-k correlation is important. We mention that for values greater than $k = 250$, a somewhat consistent precision is observed.

Altogether, it appears that centrality measures are able to explain only roughly half of the nodes that are frequently used by humans to traverse the graph. This may lead us to believe that either humans are able to assess half of the central nodes in the graph, or that existing centrality measures are simply not able to produce the portion of nodes which is considered useful by the user. In the latter case, the only remaining question is then whether or not the set of nodes returned by the centrality measures is better or worse at ensuring that a large portion of the graph is easily reachable and thus useful for completing navigation goals. We will try to answer this question in the next section.

Table I
Comparison of centrality measures for $k = 100$

	User-defined top-k precision	User-defined top-k correlation
User-defined	1.00	1.00
PageRank	0.51	0.76
Closeness	0.49	0.53
Indegree	0.37	0.83
Betweenness	0.32	0.71
HITS	0.28	0.62

B. Subgraph centrality

The final question which we aim to answer in this paper, is whether or not the top-k frequent nodes from the user-defined centrality measure are actually better or worse than graphs derived from traditional centrality measures in terms of being able to quickly reach a large portion of the original graph, and thus ensuring ease of navigation. To do this, we introduce the measure of *subgraph centrality*, which we define as the centrality (according to some existing measure, in our case closeness centrality) of a *set* of nodes, namely the set of top-k nodes obtained through a centrality measure. To determine the centrality of this set of nodes, we merge the set of top-k frequent nodes into one node, essentially realizing the equivalent of setting the weight of all edges between frequent nodes to zero.

In Figure 5 we show for increasing k the subgraph centrality values derived from the frequent nodes in the

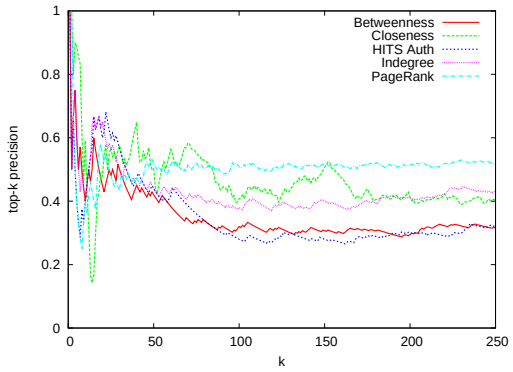

Figure 4. User-defined top-k precision (vertical axis) for different k (horizontal axis).

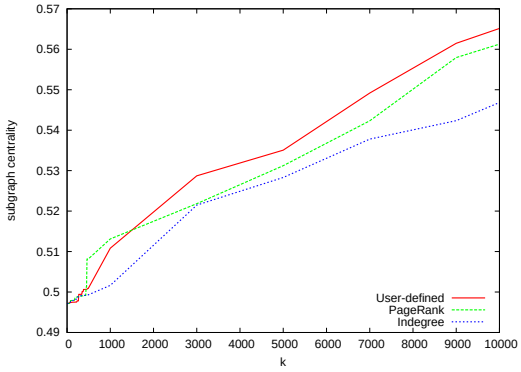

Figure 5. Comparison of subgraph centrality (vertical axis) of various centrality measures for different values of k (horizontal axis).

user-defined measure and the PageRank centrality measure. We have chosen to provide a comparison with PageRank and indegree here, because they performed best in terms of precision and correlation according to our experiments in Section V-A. We observe how the subgraph centrality of the user-defined frequent traversal graph compares quite well with that of the PageRank subgraph, which indicates that the user is able to select a portion of nodes which in terms of reachability is equal to that of a centrality measure. For $k > 1200$, the quality of the user-defined centrality is even higher than that of PageRank, suggesting that users are able to select a portion of the nodes of the graph which is better for realizing a low average node-to-node distance than a traditional measure such as PageRank.

VI. Conclusion

Throughout this paper we have looked at mining path traversal patterns from the information network Wikipedia, aiming to understand and measure the quality in terms of navigation of user-generated traversal patterns. Using data gathered from over seven millions clicks made in the Wiki Game, we have derived a new measure of node centrality based on frequently traversed nodes.

It turns out that roughly half of the set of most frequently traversed nodes overlaps with the set of central nodes according to centrality measures such as PageRank. The additional nodes that are frequently visited by the users do appear to be useful, which we have demonstrated by using the notion of subgraph centrality. Indeed, the subgraphs that can be derived from the frequently traversed nodes appear to be more central than the set of nodes derived from an existing centrality measure. This shows how users are apparently able to select an efficient portion of the graph that is useful in traversing the graph, specifically realizing a short distance to all other nodes in the graph. Although we have shown that the user is able to select an efficient subset of the graph for completing navigation goals, it remains an open question exactly *how* the user selected this subset. Clearly, a subset derived using a centrality measure or a random subset performs similar or worse, so from an artifical intelligence point of view, the performance of the user is quite remarkable.

In future work we would like to extend the study of traversal patterns to more complex patterns based on frequent edges or frequent (interleaved) subpaths. Last but not least, we plan to extend this research to other types of graphs such as social networks, in which frequently traversed nodes and edges may indicate important actors and ties in the network.

Acknowledgment

The authors would like to thank the anonymous reviewers for their useful comments and A. Clemesha for providing the clickpath data. This research is part of the COMPASS project, financed by NWO under grant number 612.065.926.

References

[1] J. Teevan, C. Alvarado, M. Ackerman, and D. 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*, 2004, pp. 415–422.

[2] B. He, M. Patel, Z. Zhang, and K. Chang, "Accessing the deep web," *Communications of the ACM*, vol. 50, pp. 94–101, 2007.

[3] A. N. Langville and C. D. Meyer, *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press, 2006.

[4] U. Brandes, "A faster algorithm for betweenness centrality," *Journal of Mathematical Sociology*, vol. 25, pp. 163–177, 2001.

[5] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "DBpedia: A nucleus for a web of open data," in *Proceedings of the 6th International Semantic Web Conference*, 2007, pp. 722–735.

[6] J. Kleinberg, "The small-world phenomenon: An algorithm perspective," in *Proceedings of the 32nd ACM Symposium on Theory of Computing (STOC)*, 2000, pp. 163–170.

[7] F. W. Takes and W. A. Kosters, "The difficulty of path traversal in information networks," in *Proceedings of the 4th International Conference on Knowledge Discovery and Information Retrieval (KDIR)*, 2012, pp. 138–144.

[8] H.-F. Li, S.-Y. Lee, and M.-K. Shan, "DSM-TKP: Mining top-k path traversal patterns over web click-streams," in *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, 2005, pp. 326–329.

[9] ——, "On mining webclick streams for path traversal patterns," in *Proceedings of the 13th ACM International World Wide Web Conference (WWW)*, 2004, pp. 404–405.

[10] R. Agarwal, K. Veer Arya, and S. Shekhar, "An architectural framework for web information retrieval based on user's navigational pattern," in *Proceedings of the International Conference on Industrial and Information Systems (ICIIS)*, 2010, pp. 195–200.

[11] R. West and J. Leskovec, "Automatic versus human navigation in information networks," in *Proceedings of the AAAI International Conference on Weblogs and Social Media (ICWSM)*, 2012, pp. 362–369.

[12] ——, "Human wayfinding in information networks," in *Proceedings of the 21st ACM International World Wide Web Conference (WWW)*, 2012, pp. 619–628.

[13] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, pp. 604–632, 1999.

[14] S. P. Borgatti, K. M. Carley, and D. Krackhardt, "On the robustness of centrality measures under conditions of imperfect data," *Social Networks*, vol. 28, pp. 124–136, 2006.