Analyzing User Click Paths in a Wikipedia Navigation Game

Denis Helic
Knowledge Management Institute
Graz University of Technology
Graz, Austria
Email: dhelic@tugraz.at

Abstract—Due to the enormous success of Web search technology navigation became only a second-class information seeking strategy on the Web. However, numerous studies highlight the importance of navigation as an alternative information retrieval technique to search. These studies provide evidences that the most efficient information finding occurs in the settings where search and navigation seamlessly integrate and complement each other. Recently, the research community has also recognized the importance of understanding the human navigation behavior since the knowledge on how users navigate helps in designing optimal navigation structures. In this paper we try to gain more insight in how users navigate towards a known target page in Wikipedia. To that end, we conduct an initial analysis of user click paths from a Wikipedia navigation game. In addition, we compare the structure of Wikipedia navigational paths with the structure of search paths in social networks and routing paths in general complex networks.

I. INTRODUCTION

In the hypertext systems of the first generation navigation was the primary method of retrieving information. The Web, as a special type of a hypertext system, adopted navigation from its very beginning [1]. Nonetheless, whereas in traditional hypertext systems the content providers always exercised the complete control over the system navigational structures making it possible for them to design efficient and optimal navigation structures, on the Web – because of its technological and social properties – the content providers have never possessed that control. The practice of Web links creation has always been governed by both the underlaying technology and social processes occurring on the Web such as e.g. preferential attachment [2]. As a result, the navigational structures on the Web emerge in a decentralized manner. Moreover, in the early days of the Web these decentralized navigational entities exhibited exponential growth and thus users very soon were not able to handle neither their sheer size nor their complexity (cf. the discussion by Jacob Nielsen on this topic [3]). Therefore, the focus of the Web developers moved to search technology as an alternative information retrieval method.

However, from the information retrieval perspective navigation represents a complementary information seeking strategy, which is for certain kind of problems much more efficient than keyword-based search. For example, in cases where users do not know about the existence of a relevant document or can not formulate their information need in a couple of keywords [4] navigation is preferred and more efficient than search. Some other researchers accentuated the importance of navigation and the complementary nature of navigation and search. For example, Teevan [5] suggested that the participants of their study on the user behavior in locating information on the Web commonly employed so-called direct situated navigation, or orienteering. In orienteering, information is reached through a large first step, e.g. a search query and then through a number of smaller steps by means of browsing. A recent study that investigated user behavior in Web search [6] showed that only few users satisfy their information need with their first search query. Instead, users visit one of the first search results, follow links on that result page, backtrack, follow some other links, then frequently refine their search, and so on.

A recent large-scale study on user behavior on the World Wide Web also supports these results. In this study authors estimated the fraction of users following links as opposed to "teleportation" [7]. More precisely, the study empirically investigated teleportation parameter $\alpha$ [8] that was introduced by Page and Brin in their PageRank algorithm. The teleportation parameter models the probability that users do not follow a link from a currently displayed document, but that they rather "teleport" themselves to an arbitrary document by e.g. typing a new URL into the browser address bar or by typing a keyword into the browser search toolbar. The original Google design assumed $\alpha = 0.15$ due to computational practicability but also due to the expectation that teleportation on the web takes place at this rate – (in other words: Page and Brin estimated that users follow links in 85% of cases). The empirical analysis of the teleportation parameter [7] shows a slightly different picture in navigational behavior of modern Web users – on the general Web, users on average follow links in around 65% of cases. Although this number is smaller then the original assumption by Page and Brin, the findings still provide strong evidence for the significance of navigation on the Web.

Although the importance of navigation seems indisputable we, as a research community, know very little about how users navigate. There are some initial empirical and theoretical studies on user navigation behavior (we mention some of these in Section II). However, many questions regarding navigation remain unanswered, such as (i) what are the mechanisms that govern user navigation in e.g. decentralized information networks, (ii) what are the structural patterns in user navigational
paths, (iii) what strategies users apply when they navigate, (iv) how does the underlying network structure influences the strategies, patterns, and navigation mechanisms, to mention only a few.

In this paper we try to provide some initial thoughts on how to answer those questions. User navigational behavior is a complex process that is dependent on various factors such as user background knowledge, current information need, information and navigation structures provided by the system, network structure, and so on. The goal of this paper is to make information and navigation structures provided by the system, such as user background knowledge, current information need, is a complex process that is dependent on various factors how to answer those questions. User navigational behavior (iv) how does the underlying network structure influences the paths, (iii) what strategies users apply when they navigate, To that end, we analyze click paths of users who participated in a small initial step and shed light on some of these factors. To that end, we analyze click paths of users who participated in a Wikipedia navigation game\(^1\). Thereby, the focus of our analysis is on the structure of these click paths.

II. RELATED WORK

One of the first studies that investigated the user navigational behavior on the Web was conducted by Huberman et al. [9]. In that study, the authors introduced a model of user navigation that assumes that users navigate to yet another Web page as long as the value of the next page exceeds a certain threshold. The result of this model is a probability distribution of the number of pages that a user visits on a Web site. The empirically obtained probability distribution follows a power law where the majority of users visits only a few Web pages and only a few users visit many pages. Thus, the value of navigation to a new page decreases as a power law with each new page that users visits.

Along the lines of these results, Pirolli developed the notion of information scents [10] and information foraging theory [11]. An information scent is an indicator of the value of each new page that a user might visit. The theory states that as long as information scent of a page is high the users are willing to explore that page. Information scent varies not only from a user to a user but also from one information retrieval task to another. For instance, if a user aims to reach a Web page on relativity theory then Albert Einstein’s Wikipedia page possesses a high information scent, however if the user wants to reach homepage of a soccer club in England then the information scent of the page on Einstein is rather low. Thus, according to information foraging theory user behavior in information systems is guided by a constant estimation of the cost and the value of a particular information item in respect to the current information retrieval task.

Research on navigation in complex networks was initiated by the problem of search in social networks, which in turn originates in the famous small-world experiment conducted by Milgram [12]. In that experiment randomly selected persons from Nebraska got a letter for a target person – a stockbroker in Boston. The experiment participants were then instructed to pass the letter to that target person through their social networks, i.e. they were allowed to send the letter only to a person whom they knew on the first name basis. Additionally, they were required to justify their decisions by writing down the reasons why they selected a particular person. In most cases the search strategy was based on geography or profession, i.e. the experiment participants either sent the letter to someone whom they knew in Boston or to a person that worked in the financial industry. The striking result of the experiment was that those letter chains that reached the target person were very short – the average chain length was only six. Thus, the population of the United Stated constituted a “small world”.

Apart from the findings that humans in the US social network are connected by short paths, another important result of that experiment was the conclusion that humans are able to find such short paths with only local knowledge of the network, i.e. humans can efficiently perform decentralized search. Kleinberg concluded that social networks possess certain latent structural properties known to humans. This background knowledge of the network structure allows us to find a short path to an arbitrary person [13], [14], [15] even in the absence of global knowledge. Basically, Kleinberg argued that humans utilize the notion of similarity or distance between nodes in a network and that they base their search strategies on that similarity, e.g. always selecting the most similar node to the target node as the next step in their search. Kleinberg represented such background knowledge as a hierarchy of nodes, where more similar nodes are situated closer to each other in the hierarchy. Also, Watts [16] introduced the notion of social identity as a membership in a number of social groups organized in multiple hierarchies and showed that an efficient decentralized search algorithm exists in the cases where humans consult those hierarchies while they search.

In [17] the authors extend the notion of distance as introduced by Kleinberg to a notion of hidden metric spaces which steer both navigation in the network and the network formation and emergence of network structural properties such as power-law degree distributions and high node clustering. The authors then connect some observable emergent structural properties of a network with its navigability by defining a region of navigable networks in two dimensional space with clustering-coefficient [18] and power-law exponent as dimensions. In such hidden metric spaces nodes are identified by their coordinates – distance between nodes is their geometric distance in a particular metric space. An interesting research question is the structure of such hidden metric spaces that underlie observable networks. In [19], the authors introduce a model with the circle as a hidden metric space and show its effects on routing in the global airport network. In [20] the authors discuss hyperbolic geometry as a hidden metric space whereas in [21] the authors apply hyperbolic geometry as a model of the hidden metric space of the Internet and design a novel greedy Internet routing algorithm.

III. DATASET AND EXPERIMENTAL SETUP

A. Wikipedia Network

In this paper we analyze the dataset containing click paths of users of the Wikigame. Wikigame is a navigational game where users navigate from a randomly selected Wikipedia page to another also randomly selected Wikipedia page. The only

\(^1\)http://thewigame.com/
possibility to reach the target page is by clicking links (the search field is e.g. not visible during the game). The complete Wikigame dataset contains about 1 million games – around one quarter of these games users finished successfully, i.e. they were able to reach the target page. In this paper we concentrate our analysis only on those successful click paths – we analyze the structure of these paths.

For the purpose of this analysis we also investigate the underlying decentralized Wikipedia network. To that end, we downloaded a complete Wikipedia dump of the English version from January 4th 2012\(^2\). We then constructed a network where Wikipedia pages are represented as network nodes and hyperlinks between those pages are represented as links connecting network nodes. There are various Wikipedia dump and various possibilities to construct such a Wikipedia network. For example, there exist different types of Wikipedia pages, as well as different types of hyperlinks. Page types are defined by so-called namespaces, with standard Wikipedia pages belonging to the main namespace, and media objects such as images or videos belonging to file namespace. There are namespaces for discussion of the page and media content called talk namespaces. Further, Wikipedia user pages belong to the user namespace. There is also a special namespace called category namespace where Wikipedia category pages are situated. Because the majority of the click paths that are available in the Wikigame dataset contain visits to either standard Wikipedia pages, media objects, or category pages we selected only those page types for the network construction. The final Wikipedia network contains 10 million nodes (10286335) and around 250 million links (254650201).

### B. Structural Analysis

Before we start with the analysis of click paths we perform the structural analysis of the Wikipedia network. This structural analysis contains the following steps.

#### Small-world properties.

According to Watts [18], a small-world network is a navigable network that is highly connected and in such a network each pair or almost each pair of nodes is connected by a short path. More formally, a navigable network posses a large connected component (typically called giant component) that contains a huge majority of the nodes and such a navigable network exhibits a diameter (the longest shortest path in the network) that is proportional to the logarithm of the number of network nodes:

\[
D \propto \log N. \tag{1}
\]

Alternatively, this second property is often expressed in terms of the average shortest path or in terms of the effective diameter, which is the longest shortest path under which 90\% of nodes are reachable from each other.

#### Clustering and Network Heterogeneity.

Another two network structural properties that are commonly observed in navigable networks are high clustering and degree heterogeneity. In highly clustered networks there is a large number of closed triads, i.e. in such networks the probability that two nodes are connected in cases when they share a common neighbor is higher than in e.g. random networks. For example, social networks are typically highly clustered as the probability that a friend of a friend is also a friend is high since humans tend to connect with people similar to themselves. It has been shown [18], [22], [17], [19] that such high clustering combined with a few so-called long-range links is necessary for establishing network navigability.

Additionally, many real-world networks exhibit heterogeneous degree distributions. In particular, power-law distributions are very common (see e.g. [23], [24], [25], [26]). Such networks have a few of highly connected nodes called hubs and many nodes (the long tail) with only a few links. In [27] the authors analyze and show the importance of hubs for navigation. In particular, the authors present an algorithm for searching in power-law networks that in the opening phase of search applies a simple strategy – the algorithm proceeds along increasing degrees and reaches very quickly one of the hubs. As hubs have many connections the probability of reaching the destination node from a hub is higher than reaching that destination node from a low degree node. The authors also analyze the time complexity of such an algorithm and show that the time needed to reach the destination node is sub-linear.

In [19] the authors establish a connection between the power-law exponent, clustering coefficient and the network navigability. By simulating navigation in networks generated with different values of power-law exponent and clustering coefficient the authors identify a region of navigable networks in this two dimensional space. Additionally, they show that many real-world networks are efficiently navigable as the observed values of the power-law exponent and the clustering coefficient are situated within the navigable region.

Apart from in-degree \(k_{\text{in}}\) and out-degree \(k_{\text{out}}\) distributions we also analyze the \(k_{\text{in}} / k_{\text{out}}\) distribution. This ratio captures the “flow” of links towards and outwards of a node and can be a better estimator of the node importance for navigation. In [28] the authors apply this ratio among other measures to quantify the hierarchical level of nodes. Although very simple this ratio performs good and outperforms some other more complex measures such as PageRank [8] or HITS [29].

### C. Click Path Analysis

In the second part of our analysis we turn to the click paths. This analysis phase contains the following steps.

#### Distribution of Click Path Lengths.

First, we analyze click path length distribution. We are interested here in the ability of users to find short paths in the network in the case when those paths exists.

#### Degree Distributions of Visited Nodes.

We start the click path structure analysis by investigating the degree distributions of visited nodes. With this analysis we shed light on the relative importance of nodes of different degrees for the user navigation. We also compare these distributions with the results from similar studies dealing with search in other types

\(^2\)http://dumps.wikimedia.org/enwiki/
Fig. 1. Small-world network properties in Wikipedia network. The network is highly connected – the largest strongly connected component includes around 55% of nodes (left). If we ignore the link direction the largest weakly connected components contains more than 99% of nodes (center). Wikipedia nodes are connected by short paths – the effective diameter is less than 6 (right).

Fig. 2. Degree distributions in Wikipedia network. Degree distributions follow, at least in certain regions, power-law. Nevertheless, the distributions are clearly skewed and heterogeneous. They show the existence of the hubs and the long tail of nodes with low degrees. In the right figure we plot \( \frac{k_{in}}{k_{out}} + 1 \) distribution to correct for division with zero and plotting on the log scale.

Fig. 3. Distribution of click path lengths. The distribution is similar to e.g. original small-world experiment [12]. This can be an indicator that user navigation in information networks follows similar principles as human search in social networks. Users are typically very efficient and can find short paths quickly. The average click path length is \( \bar{l} = 6.27 \) with standard deviation \( \sigma = 2.42 \).

of networks, e.g. [27], [12], [19]. Finally, we also analyze the \( \frac{k_{in}}{k_{out}} \) distribution.

**Navigation Patterns.** In [19] the authors observed a navigational pattern in complex networks that consists of two phases. In the first phase search starts at a low degree node in the network periphery and from their continues by “zooming out” from the periphery to the network core. Once within the network core the navigation proceeds to a hub, where a transition to the second “zoom in” phase occurs. In this second phase the navigation continues out of the network core to the final destination, again typically in the network periphery. The phases can be recognized by constantly decreasing distances to the destination node and increasing degrees of visited nodes in the “zoom out” phase followed by decreasing degrees in the “zoom in” phase after the transition in the network core occurs. In this part of our analysis we also try to identify similar pattern in user navigation click paths.

**IV. RESULTS**

**A. Structural Analysis**

Wikipedia network clearly exhibits small-world properties (see Figure 1). The effective diameter of the network is slightly less than 6, which means that the huge majority of the nodes is reachable from each other in only a few clicks. The global network topology is dominated by a giant component, i.e. the largest strongly connected component includes more than 55% of nodes. This topology resembles the topology of the Web that is best described as the bow-tie model [32]. In this model, this strongly connected giant component occupies the network core and is connected to the IN component via back-links and to the OUT component via forward links. There are also direct connections between IN and OUT called TENDRILS. In an undirected case the weakly connected component contains all four elements SCC, IN, OUT, and TENDRILS. The size of the weakly connected component is around 99%. The
connectedness of Wikipedia is higher than that of the Web – on the Web all four components contain approximately one quarter of all nodes. Please note that our results are different than e.g. [33] as the size of the strongly connected component in that study was around 85%. The reason for these different results are different approaches to the Wikipedia network construction. In [33] the authors included also user, help, and talk pages into the network and this influences the network connectivity.

**Clustering and Network Heterogeneity.**

Degree distributions in Wikipedia network are heterogeneous and skewed (see Figure 2). In certain degree regions both in-degree distribution ($k_{in}$) and out-degree distribution ($k_{out}$) follow power-laws. Nevertheless, the existence of hubs and long tails in those distribution is clearly visible. Using the estimation method introduced in [34] we estimate the power-law exponent and the minimal degree of the power-law distributions. We obtain $\gamma = 2.28$ and $k_{min} = 1,668$ for in-degree distribution and $\gamma = 3.21$ and $k_{min} = 254$ for out-degree distribution. Please note that these results are comparable with other studies that estimated power-law exponents in Wikipedia networks, e.g. [33]. The $k_{min}$ exhibits also power-law behavior.

The average clustering coefficient of the Wikipedia network is around 0.33 which is very high. According to [19] a network with such combination of the values for clustering coefficient and power-law exponents is a navigable network.

**B. Click Path Analysis**

**Distribution of Click Path Lengths.**

The click path length distribution of successfully finished games is shown in Figure 3. A simple conclusion here is that whenever users are able to find the destination node they are typically extremely efficient in doing so. The average click path length is slightly higher than the effective diameter of the network – 6.27 as compared to 5.70 – meaning that on the average users need only a half click more than the global shortest path. Also, the shape of the click path distribution resembles the shape of the distribution of search path lengths in the original small-world paper by Milgram [12]. Both of these findings can be a strong indicator that user navigational strategies are similar to those strategies that humans apply for the search in their social networks.

**Degree Distributions of Visited Nodes.**

Now we turn to the analysis of the importance of the nodes of different degrees for the navigation process. For
that purpose we analyze the distributions of in-degree, out-degree and the ratio of in-degree and out-degree of nodes visited during navigation. All of the distributions show the existence of certain specific nodes, typically with high degrees or high degree ratio, which are visited in the large fraction of click paths. These nodes seem to be extremely important for navigation – they are navigational hubs (see Figure 4).

Navigation Patterns.

For in-degrees and in-degree/out-degree ratio we observe similar two phase navigational patterns as reported by [19] in our dataset. Thus, in the initial phase of their game users reduce the distance to the destination node by visiting nodes of increasing degrees, i.e. users enter the network core. In the network core transition to the second navigation phase occurs – users navigate to nodes in the periphery selecting the most similar nodes to the destination node. This phase is characterized by a decreasing distance to the destination and decreasing degrees of the visited nodes. We analyze this pattern by the density maps shown in Figure 5. Out-degrees seem less important for establishing navigational patterns.

This analysis also shows the importance of high degree nodes, as well as nodes with high degree ratio for the navigation. Typically, transition between two navigation phases occurs on such navigational hubs.

V. CONCLUSIONS

In this paper we presented the results of the initial analysis of user click paths in a Wikipedia navigational game. Our main contributions can be summarized as follows:

- Users are very efficient at navigation. They are able to find short paths between randomly selected pages.
- User navigation is a two phase process – in the first phase users “zoom out” to the network core and in the second phase they “zoom in” to their final destination.
- Certain high in-degree nodes and nodes with high “flow” ratio are very important for navigation. These nodes are visited very frequently – these are navigational hubs.

REFERENCES