Human Wayfinding in Information Networks

Robert West
Computer Science Department
Stanford University
west@cs.stanford.edu

Jure Leskovec
Computer Science Department
Stanford University
jure@cs.stanford.edu

ABSTRACT

Navigating information spaces is an essential part of our everyday lives, and in order to design efficient and user-friendly information systems, it is important to understand how humans navigate and find the information they are looking for. We perform a large-scale study of human wayfinding, in which, given a network of links between the concepts of Wikipedia, people play a game of finding a short path from a given start to a given target concept by following hyperlinks. What distinguishes our setup from other studies of human Web-browsing behavior is that in our case people navigate a graph of connections between concepts, and that the exact goal of the navigation is known ahead of time. We study more than 30,000 goal-directed human search paths and identify strategies people use when navigating information spaces. We find that human wayfinding, while mostly very efficient, differs from shortest paths in characteristic ways. Most subjects navigate through high-degree hubs in the early phase, while their search is guided by content features thereafter. We also observe a trade-off between simplicity and efficiency: conceptually simple solutions are more common but tend to be less efficient than more complex ones. Finally, we consider the task of predicting the target a user is trying to reach. We design a model and an efficient learning algorithm. Such predictive models of human wayfinding can be applied in intelligent browsing interfaces.

Categories and Subject Descriptors: H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia—*Navigation*.

General Terms: Algorithms, Experimentation, Human Factors.

Keywords: Navigation, browsing, information networks, Wikipedia, Wikispeedia, human computation.

1. INTRODUCTION

There is no such thing as an isolated piece of knowledge. Bits of information are interconnected in giant networks, and we are daily navigating and finding paths through such networks. Browsing the Web is an important example, but by far not the only one: we follow leads in citation networks to find work that is related to our own research; when we reason or try to find explanations for the phenomena around us, we are implicitly disentangling a network of relations between concepts, with the goal of finding a path of connections between the 'cause' and the 'effect'; and we constantly look things up in cross-referenced dictionaries and encyclopedias, be it in the form of books or online resources such as Wikipedia.
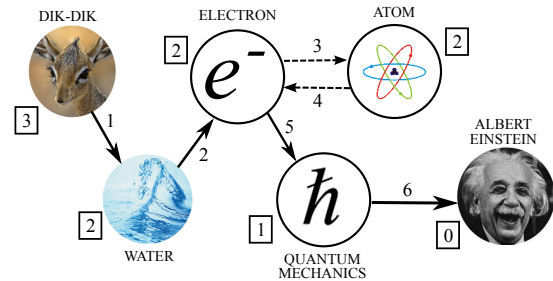
Figure 1: A human example path between the concepts DIK-DIK and ALBERT EINSTEIN. Nodes represent Wikipedia articles and edges the hyperlinks clicked by the human. Edge labels indicate the order of clicks, the framed numbers the shortest-path length to the target. One of several optimal solutions would be ⟨DIK-DIK, WATER, GERMANY, ALBERT EINSTEIN⟩.

This last example is particularly interesting, since Wikipedia is not just a regular website but a rich network representing human knowledge as well as the connections between single pieces of knowledge, by means of hyperlinks. This distinguishes Wikipedia-browsing from navigation on typical Web resources. By observing humans finding their ways between articles in Wikipedia, we are watching them navigate a large information network, using their mental maps of relationships in order to find the paths that connect concepts.

There are two aspects—analytic and pragmatic—of this view of human wayfinding in information networks.

From an analytic perspective, it is important to understand what strategies and clues people use to find paths in the Wikipedia information network. In particular, as humans are navigating information networks, they might switch between various strategies. The interplay between the topical relatedness of concepts and the underlying network structure could give us important insights about the methods used by efficient information seekers. Also, the latter often face trade-offs: there may be wayfinding strategies that are safe but also inefficient; on the other hand, by trying to find only the shortest paths, the searcher might get lost more easily.

From a pragmatic perspective, there is useful information in the trail an information seeker has navigated so far, even before reaching the target. We see such trails playing an important role in the development of methods that can analyze the path taken so far and provide information seekers with navigational aids. One useful direction for this is in predicting what piece of information the information seeker is trying to locate. Another is in automatically detecting if the user has gotten lost. Given that human navigation of information networks is so ubiquitous, a better understanding of the methods according to which humans find connecting paths

would have applications in improving the design of information spaces [15], more intuitive and navigable link structures [8], and new intelligent information navigation systems [16].

Present work. These broad issues suggest a wide range of interesting open questions. We take a step in this direction by computationally analyzing how people navigate to specific target pages in the Wikipedia information network. As a tool we use the online human-computation game Wikispeedia [24, 23], in which players (i.e., information seekers) are given two random articles and aim to solve the task of navigating from one to the other by clicking as few hyperlinks as possible. Players have no knowledge of the global network structure but must rely solely on the local information they see on each page—the outgoing links connecting the current article to its neighbors—and on their expectations about which articles are likely to be interlinked. In this respect, the task humans are trying to solve at each visited article is that of guessing which of the outgoing links to follow in order to eventually reach the target article.

What makes our study unique is that we have been collecting detailed data on more than 30,000 instances of human wayfinding in an information network describing general human knowledge (the data came from around 9,400 distinct IP addresses). This allows us to computationally analyze human wayfinding on a large scale. Even more important, for every instance we know the starting article and the given target article the user is trying to reach. Hence, we do not have to infer or guess the information need of the information seeker, but can base our methods on the ground truth instead.

To illustrate the dynamics of the Wikispeedia game, as well as potential reasoning schemes and classes of strategies humans might use, Fig. 1 gives the example of a human path between the *start article* DIK-DIK and the *target* ALBERT EINSTEIN. (We call such a pair a *mission.*) Note that using the browser's back button is allowed. In the example, the information seeker clicked from ELECTRON to ATOM, but backed up after not finding the link to the target that he/she had expected there. We call the sequence including ATOM and the *back-click* the *full path*, while referring to ⟨DIK-DIK, WATER, ELECTRON, QUANTUM MECHANICS, ALBERT EINSTEIN⟩ as the *effective path*. The *shortest-path length* (SPL) from every article to the target is shown in squares in the picture. If a click decreases the SPL, we call it *lucrative*. Note that, in the example, not every click is lucrative; rather, the information seeker makes progress at first, but then orbits at a distance of 2 from ALBERT EINSTEIN, before finally gravitating towards it with the choice of QUANTUM MECHANICS. We also emphasize the special role the article on WATER plays in the example. It connects to many parts of the network—hence we call it a *hub*—and marks the transition between getting away from the animal kingdom and homing in on the realm of physics.

Despite the lack of global knowledge, humans are good at connecting the dots: the median human game path is only one click longer than the median optimal solution. We explain this effect by showing that certain properties of Wikipedia's hyperlink structure make it easily navigable. For instance, our Wikipedia graph (we use a version containing about 4,000 articles and 120,000 links [27]) has a skewed degree distribution (median/mean/max degree 19/26/294) and contains a few high-degree hubs that contribute to everything being connected to everything else by short chains (median/mean/max shortest-path length 3/3.2/9; note that this is the case although no 'meta pages', such as category indices, are available to players). This makes our network a typical 'small world'. Our analysis shows that people commonly find their way in it by first locating a hub and by constantly decreasing the conceptual distance

to the target thereafter. While approaching the target through a series of conceptually very related articles is safer and often humans' preferred solution (cf. the example of Fig. 1), it is typically not the most efficient: we find that thinking 'out of the box' often allows information seekers to find shorter paths between concepts—at the risk of getting lost. A strategy that is both popular and often successful is to connect concepts in terms of their geographical commonalities. In the above example, ⟨DIK-DIK, AFRICA, EUROPE, GERMANY, ALBERT EINSTEIN⟩ would have been such a solution.

Following this analysis, we formulate a task that captures some of the key motivating issues discussed above. We show how information from a short prefix of the navigation path can be used to predict what the information seeker is looking for. We design a ranking-based machine learning model and an efficient parameter estimation algorithm. Our method is informed by the lessons learned in our analysis and is trained on real human paths. The experimental evaluation shows that it can predict humans' intended targets with high accuracy.

Overall, our results provide insights into how people navigate and solve the task of wayfinding in information networks. From the practical perspective, our findings can be applied in order to make better sense of observed human search paths. Our performance on the target prediction task suggests that features of the underlying path can provide useful information beyond simply predicting the next action of the user. We therefore think that results of our research can be incorporated into intelligent systems to facilitate human information browsing and navigation.

2. RELATED WORK

The work related to our explorations here can be separated into three parts: Web click-trail analysis, systems that aid users in Web navigation, and decentralized search in networks. Next, we briefly review each of these three lines of related work.

Information retrieval has focused on analyzing Web-browsing click trails of millions of users mainly for the purpose of improving Web search results. Click trails can be used as endorsements to rank search results more effectively [4, 20], trail destination pages can themselves be used as search results [26], and the concept of teleportation can be used to navigate directly to the desired page [21]. Similarly, large-scale studies of Web-page revisitation patterns [2] focus on how often users revisit the same page, while ignoring how people get there. In contrast, our work focuses on understanding how people reach information by navigating through networks. Another important difference is that, in our case, we know the exact target of human search and can thus quantitatively analyze the strategies people use when navigating information spaces, as well as their efficiency.

Observational and laboratory studies have conducted small-scale controlled experiments about users' thought processes during Web search by having them think aloud as they search [14], and about their interaction with Web information [18]. These studies spawned sophisticated descriptive models, like information scent [5] and information foraging [17], which uses the metaphor of how animals forage for food. Other analogies, such as orienteering [15] and berrypicking [3], have also been used to describe users' information-seeking strategies. Systems like ScentTrails [16] and guided tours [22] have been proposed to create annotations to indicate where other users have navigated in the past, all with the goal of helping people find information faster. Our present work differs in two important ways: First, our goal is not to formulate an analogy for human wayfinding, but rather to analyze it computationally using a large-scale collection of real search traces. Second, we ad-
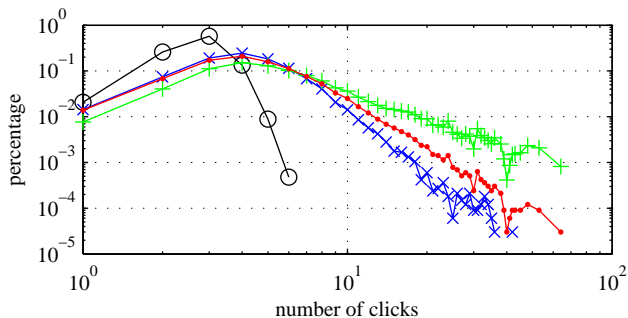
Figure 2: Distribution of game length, according to different path-length metrics. *Black circles:* **shortest possible paths.** *Blue X's:* **effective human paths (i.e., ignoring back-clicks).** *Red dots:* **complete human paths (i.e., including back-clicks).** *Green plus signs:* **complete human paths, corrected for drop-out rates.**

path-length metric	mode	median	mean
shortest possible paths	3	3	2.9
human, effective	4	4	4.9
human, incl. back-clicks	4	5	5.8
human, drop-out–corrected	4	6	8.9

Table 1: Summary statistics of the distributions of Fig. 2.

dress the task of predicting the actual target of human search, not just the next action [7].

The last line of related work can be traced back to Milgram's small-world experiment [13] and the algorithmic problem of decentralized search in networks [10]. Decentralized search considers a scenario in which a starting node s is trying to send a message to a given target node t by forwarding the message to one of its neighbors, where the process continues in the same way until eventually t is reached. This process has been investigated both experimentally as well as through simulations [6, 11, 1, 9, 19]. Each game of Wikispeedia may be considered an instance of decentralized search in a network, where players try to navigate between given start and target pages using only the local information provided on the current page (i.e., players can only follow hyperlinks of the current page). In the small-world experiment, search is in a sense even more decentralized, since each node—i.e., human—on the path independently forwards the message and then forfeits control. While in Wikispeedia the information seeker also has only local knowledge about the unknown network, he/she stays in control all the way and can thus form more elaborate strategies than in the multiperson scenario. Moreover, as previous empirical studies of search behavior had very few completed paths (e.g., only 384 [6]), we work with more than 30,000 completed chains.

Our study is unique in several respects. We collected large-scale data about human navigation in a network of real-world concepts, where we know the precise target node humans are trying to reach. We focus on computationally investigating and modeling how humans navigate information networks and what strategies they use. This allows us to build accurate predictive models of where the users are trying to navigate.

3. EFFICIENCY OF HUMAN SEARCH

The Wikipedia graph is an example of a 'small world' in which most pairs of nodes are connected by short chains, with a mean/median/max shortest-path length (SPL) across all pairs of 3.2/3/9. A natural first question to ask is, How good are humans at finding such short chains?
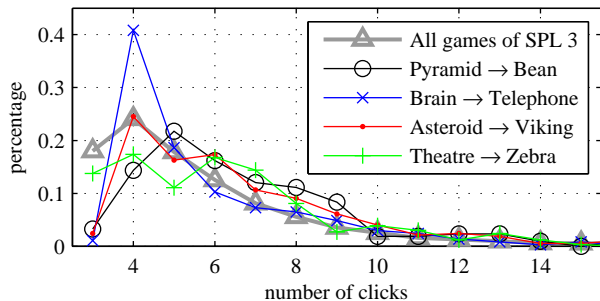

Figure 3: Distribution of game length for four specific missions with an optimal solution of 3 clicks. We recorded between 216 and 376 paths per mission. The gray curve shows the length distribution for all games with an optimal solution of 3 clicks.

Fig. 2 gives a good impression of how the paths found by humans compare to optimal solutions (summary statistics of the distributions in the figure are provided in Table 1). The red line shows the distribution of human path lengths (where clicks that were later undone and back-clicks are counted as regular clicks), while effective paths were used for the blue line. For each human game we also computed an optimal solution, and the resulting path length distribution is plotted as a black line. We make three observations:

1. The variance in search time is much larger for human than for optimal solutions. While the distribution of optimal path length is tight around 3 clicks, the human distribution exposes a heavy tail.

2. Nonetheless, the effective paths found by humans (the blue line in Fig. 2) are typically not much longer than shortest paths. Both mode and median search times differ from optimal by just 1 click (3 vs. 4 clicks), mean search time by 2 clicks (2.9 vs. 4.9 clicks). (See Table 1.)

3. When considering full path length with undone and back-clicks (the red line in Fig. 2), the mode search time is still 4, and the mean and median search times are 1 click more than for effective paths (5 vs. 4, and 5.8 vs. 4.9 clicks). That is, humans click back on average once every other game.

Two questions arise: First, what is the reason for the large variance in human search time? Second, why is human search still so efficient on average?

The first question permits two potential answers. Either some missions are inherently harder than others, or some information seekers are better than others. Some missions have longer optimal solutions than others, so necessarily some games are inherently harder. However, even when restricting ourselves to missions of a fixed SPL, the numbers stay virtually unchanged (e.g., for games with a SPL of 3 clicks, the mode/mean/median is 4/5/6.0, as opposed to 4/5/5.8 for all games). Of course, even among missions of a fixed SPL, some are harder for humans because the lucrative links might be less obvious. To control such effects, we posted four missions—all of SPL 3—on the game website with increased frequency. This allows us to find out how different humans perform on the exact same task. The search time distributions for the four frequent missions are plotted in Fig. 3. We see that for each separate mission there is considerable search time variance, but also that some missions allow for shorter games on average than others. This leads us to conclude that both hardness of mission and individual skill play a roll in explaining the large search time variance.
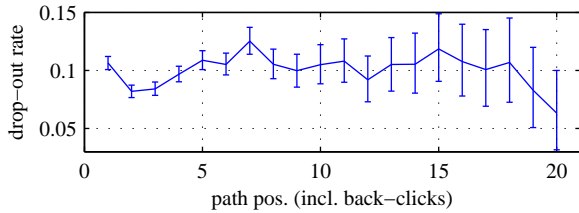
Figure 4: Drop-out rate as a function of path position (with 95% confidence intervals). At each step, players give up with a probability of around 10%.

Figure 5: Link probability $P(r)$ as a function of rank r. Given r, consider all node pairs (u, v) such that v is the node that is r-th closest to u among all nodes. Then $P(r)$ is defined as the fraction of these nodes for which u links to v. Blue: $P(r)$. Red: $P(r) + \epsilon$, with $\epsilon = 0.005$. Black: ideal slope of -1 (not a fit; only for orientation).

Regarding the second question, too,—Why is human search so efficient on average?—several answers are conceivable. One might argue that the efficiency of observed games is caused by a sampling bias. In studies that collect data from human volunteers, one always faces the problem of participants dropping out before finishing the task assigned. In our case, this might result in a bias towards observing shorter chains than what we would observe by forcing participants to finish all tasks, since the longer the game takes, the more likely the subject is to give up at some point. For instance, 54% of all games in our data set were canceled before finishing. Fig. 4 shows that the drop-out rate R_i, i.e., the probability of giving up at the i-th step, is roughly constant at around 10%.

Using drop-out rates, we can correct for the aforementioned bias and compute an ideal search time histogram, for the hypothetical case that participants never give up [6]. The result is shown as the green line in Fig. 2. Although longer games are more frequent under the ideal than under the observed distribution, the distributions still look similar qualitatively, with mode 4 and a power-law–like tail. The median search time is only 1 click higher (6 vs. 5 clicks), and mean search time rises by 3 clicks (8.9 vs. 5.8 clicks). We conclude that the observed human efficiency in Wikispeedia play is not explicable by a sampling bias alone.

Instead, we conjecture that, even without knowing the set of all existing links, the Wikipedia graph is efficiently navigable for humans because they have an intuition about what links to expect. Clearly, the probability of two articles linking to each other is higher the more related they are. This can lead to efficient navigation even in the absence of global knowledge. In particular, Liben-Nowell *et al.* [11] have shown analytically that short search times (technically defined as polylogarithmic in the number of nodes) can be expected under their model of 'rank-based friendship', viz., if the probability of a node linking to its r-th closest fellow node decays as $1/r$. Intuitively, such a scenario is desirable because it constitutes an appropriate mix of many short- and a few long-range links. The latter are helpful for getting somewhat close to the target, while the former are necessary for fully reaching it.

We strive to investigate whether the Wikipedia graph satisfies rank-based friendship. Humans may tap into all their knowledge and reasoning skills during play, so it is hard to formalize their node distance measure. In the present analysis, we therefore coarsely approximate the human by a standard text-based distance measure and define the similarity of two articles as the cosine of their TF-IDF vectors [12] (and distance as one minus similarity). Fig. 5 plots the link probability $P(r)$ as a function of rank r. The black line was added to show an ideal slope of -1, as postulated by the rank-based friendship model. Note that, although $P(r)$ does not fully follow a power law, the overall slope of the curve comes close to -1, which leads us to conclude that Wikipedia is conducive to efficient navigation because its links represent an appropriate mix of long- and short-range connections across concept space.
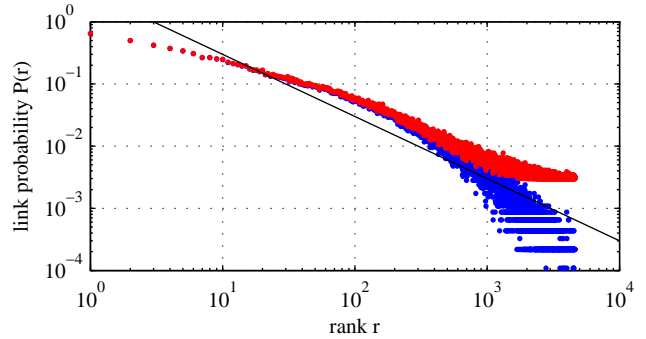
Also note the red, upper curve in Fig. 5: after adding a small constant $\epsilon = 0.005$ to $P(r)$, the plot looks considerably more like the required power law. We take this as an indication that there is slight underlinking in the Wikipedia graph: if every node linked to even its furthest fellow nodes with a small background probability ϵ, then Wikipedia could become even more easily navigable (at least under the TF-IDF distance measure).

4. ELEMENTS OF HUMAN WAYFINDING

In the previous section, we have argued that human search in the Wikipedia network is made possible by the statistical properties of its link structure. Next we turn our attention to a detailed analysis of how people actually exploit these properties.

4.1 Anatomy of typical paths

In our analysis, we investigate how some key quantities of articles and clicks change as games progress from the start towards the target article. To facilitate the analysis, we restrict ourselves to all games whose start and target articles are optimally connected by exactly 3 clicks and consider only effective paths.

Fig. 6 contains a graphical summary of the findings we are about to discuss. Each subfigure tracks one quantity along game paths; each curve is computed from all games of the same effective path length, the leftmost curve representing games of length 8, the next one games of length 7, etc. (to avoid clutter, we consider only games of a maximum length of 8 clicks). The x-axes show the human-path distance, i.e., the number of clicks to the target on the effective path (i.e., paths may be thought of as running from left to right), while the y-axes represent the mean of the respective quantity over all games, alongside 95% confidence intervals. The bold gray curves plot the given quantity for the average optimal solution. To compute it, we found an optimal solution for every human game instance and averaged. We refer to the figure in row r and column c as plot (r, c).

Making progress is easiest far from and close to the target. Plot $(1, 1)$ shows how the shortest-path length (SPL) to the target changes as a function of human-path distance. Necessarily, the shorter the game, the steeper the curve. Additionally, all curves share a typical anatomy: with the first click, the information seeker gets significantly closer to the target on average, then the curve flattens out and becomes steeper again towards the endgame. In short games, the players blasts straight through to the target, making progress with nearly every step, while in long games the player
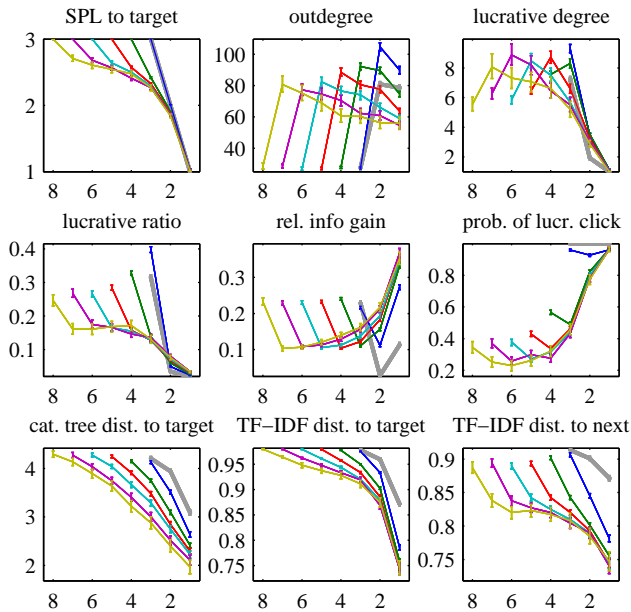
Figure 6: The evolution of article properties along search paths, for games of optimal length 3. Only games of between 3 and 8 clicks are shown. Each colored line represents games of the same length. The *x*-axis shows the distance-to-go to the target, the *y*-axis the average value of the respective property (with 95% confidence intervals). The bold gray curve is computed based on optimal solutions for the considered human paths.

goes through a phase of inefficient circling around the target before finally gravitating towards it. Another perspective of the same phenomenon is afforded by plot $(2, 3)$, which shows the fraction of times humans picked a lucrative link, i.e., one that led them closer to the goal in terms of SPL. We observe a down–up pattern in the curves: information seekers are more likely to make progress with the first click than with the second. Later on, in the endgame, clicks become again ever more likely to be lucrative. In long games, the phases of progress in the opening and endgame are separated by a phase of stagnation where the probability of picking a good link stays roughly constant, a manifestation of the circling effect described above.[1]

Hubs are crucial in the opening. The initial progress with the first click is afforded by leaping to a 'hub' article, i.e., a high-degree node that is easily reachable from all over the graph and that has connections to many regions of it. This makes sense intuitively, since a good hub gives the information seeker more options to continue the search, and is demonstrated by plots $(1, 2)$, $(1, 3)$, and $(2, 1)$. While the start article has an average degree of only about 30 (cf. plot $(1, 2)$), the first click leads to an article with an average degree of between 80 and 100. After the sudden degree increase with the first click, the quantity decreases slowly as the target is approached.

Note that the shorter the game, the higher the degree of the hub (and of any given position, for that matter). This could mean (1)

---

[1] The fact that the probability is not 100% even when humans achieve the optimal path length (the blue curve) is due to the fact that players might have later undone clicks taken from articles along the effective path by means of the browser's back button, such that they may have taken suboptimal links while still achieving the optimal effective path length.
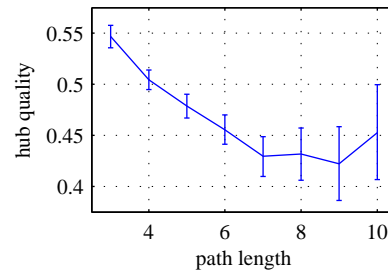

Figure 7: Hub quality as a function of search time (with 95% confidence intervals). Hub quality is defined as the degree of the second article, divided by the degree of the maximum-degree neighbor of the start article.

that better information seekers pick better hubs, or (2) that some missions are easier because the start articles have links to better hubs. While the availability of good hubs certainly helps, Fig. 7 demonstrates that the first alternative plays a role as well. We plot the ratio $\deg(u_2)/\deg(u_2^*)$ of the degree of the second article and that of the highest-degree neighbor of the start article, averaged over all games of the respective length. The quantity decreases with increasing game length, implying that better information seekers tend to start games with relatively higher-degree hubs.

Let the term 'lucrative degree' stand for the number of outgoing links that decrease the SPL to the target. Plot $(1, 3)$ shows that, just like the plain degree, the lucrative degree, too, increases significantly with the first click—the hub article typically offers more lucrative options than the start article. Also, the mean lucrative degree then decreases as the games continue (necessarily, since there are more articles far from the target than close to it). We do not see a correlation between the hub's lucrative degree and game length. However, the start article itself has higher lucrative degree for very short games than for longer ones, an indicator that some games are inherently easier than others, even if the optimal number of clicks is held fixed. This certainly is a factor in the aforementioned negative correlation between search time and hub degree: if there are many good hubs it is easier to find one of them.

An interesting additional insight is afforded by looking at how the average of the ratio of lucrative degree and degree changes during games (cf. $(2, 1)$). The resulting quantity, which we call 'lucrative ratio', corresponds to the probability of getting closer to the target when randomly choosing an outgoing link. While both degree and lucrative degree achieve their maximum with the second article, their ratio drops drastically between the first and second articles. From this we conclude that the second article is a true hub, in that it does not only have many outlinks leading closer to the target, but has even more that lead further away from it, i.e., that it has connections into many different regions of the graph.

Conceptual distance to the target decreases steadily. Plots $(3, 1)$ and $(3, 2)$ show that articles get ever more related textually to the target as the latter is approached (in other words, textual distance decreases). We verify this using two distinct measures of conceptual relatedness, (1) the cosine of the TF-IDF vectors of the two respective articles, as in Section 3, and (2) the number of edges that have to be traversed in the category tree that comes with our Wikipedia version, in order to reach one article from the other ('category tree distance'). The fact that the conceptual distance to the target decreases strictly along paths corroborates our conjecture from Section 3 that humans approximately perform a decentralized search using a distance measure between concepts. Also, note that the very intuition that the distance between concepts along the path

and targets decreases was the original *raison d'être* of the game of Wikispeedia [24].

Big leaps first, followed by smaller steps. While plots $(3,1)$ and $(3,2)$ track the textual distance between the current article and the target, plot $(3,3)$ does so for the distance between the current and the next articles. This 'textual step size' is monotonically decreasing: first, information seekers make big leaps, with adjacent articles being rather unrelated (e.g., when jumping to the hub); then, as they home in on the target, they straddle ever smaller 'gaps'. This progression is possible because Wikipedia's link structure trades off long- versus short-range connections in a favorable manner, as laid out in our discussion of rank-based friendship in Section 3. We also see the aforementioned circling effect for long games again: between the initial getting-away and the final homing-in, both the textual distance to the target and the textual step size stagnate, as the player stumbles around on the graph.

Clicks are most predictable far from and close to the target. Finally, consider plot $(2,2)$, which attempts to capture the agreement between different humans. Consider a target article t. For each article u, we define a click probability distribution over u's outlinks, which counts for each outlink how often it was taken when humans were searching for the target t (with add-0.1 smoothing, to mitigate the effect of zero counts). The entropy of this distribution provides us with a measure of how predictable human clicks are, lower entropy meaning higher predictability. We let the term 'information gain' refer to the difference between the prior entropy of the uniform click distribution before observing any clicks and the posterior entropy given all game data. It measures how much more predictable clicks at a given article u are after seeing the game data than before. 'Relative information gain' is the ratio of information gain and prior entropy, or in other words, the percentage-wise decrease in uncertainty afforded by observing the game data. This quantity exposes a characteristic pattern, as shown in plot $(2,2)$. The relative information gain at the start article is typically around 23% on average and much lower (around 10%) for the following article. The leap to the hub is much more predictable than the ways in which people continue from there. (This is compounded by the fact that, given a start article, not all humans choose the same hub, such that for each hub we have fewer samples than for the start article and the respective click distribution stays more uniform, resulting in higher posterior entropy and thus lower information gain.) As information seekers approach the target, their behavior becomes again more coherent and predictable, with information gain increasing.

Comparison of human with shortest paths. To conclude our discussion of typical human search paths, we compare them to the optimal solutions found by a shortest-path algorithm (the bold gray curve in each plot). Most of the curves are qualitatively similar to those for human paths, which follows from the structural constraints imposed by the link graph. However, there are quantitative differences with respect to all quantities we investigate. For instance, for shortest paths, too, the average degree goes up with the first click, but this is purely statistically so because the shortest-path finder is more likely to pick high–betweenness-centrality nodes, which in turn tend to have high degree; note that nonetheless the hub has about 20 fewer outlinks than for optimal humans. The lucrative degree of the hub is about 3.5 for optimal solutions found by humans, while it is and only 2 for solutions found by the shortest-path finder.

The relative information gain is nearly zero for the second article, much smaller than the 10% typical for humans. The reason is

that shortest paths are often entirely different from human paths, such that the second article itself is often one that humans never picked. Since the information gain is computed solely based on human paths, the entropy at the second article stays very uniform (i.e., information gain close to zero).

The curves for TF-IDF similarity to the target and to the next article are qualitatively similar to those for human paths, in that the distance values decrease as games progress. This is due to the fact that closeness in the Wikipedia graph is correlated with textual similarity (cf. Fig. 5). Therefore, as the graph distance to the target decreases, so does the textual distance (plots $(3,1)$ and $(3,2)$). Note, however, that the decrease is much more pronounced for human paths: humans explicitly navigate according to the content of articles, while the shortest-path finder does so only because it is implicitly constrained by the statistical properties of the hyperlink graph.

4.2 Trade-off between similarity and degree

Given the findings of the previous section, degree and similarity seem to be the most important factors in human wayfinding in Wikipedia. We hypothesize that humans navigate more strongly according to degree in the early game phase, when finding a good hub is important, and more strongly according to textual similarity later on, in the homing-in phase. The goal of this section is to test this hypothesis.

We conduct the following experiment to gauge the trade-off between similarity and degree. Consider only the games with an optimal solution of 3 clicks. Then divide the set of all human trajectories into subsets according to the number of clicks taken by the player (the maximum length we consider is 8).[2] Each of these subsets is divided into balanced training (70%) and test (30%) sets. For each training set and each path position in the training set, we train a logistic regression classifier, using two features (and a constant bias term), representing degree and similarity to the target, respectively. The positive examples consist of all human clicks contained in the respective training set. The negative examples have to be contrived (since we have no ground truth of clicks a human will never make). We do so by randomly (with replacement) sampling clicks that were never observed, until there are as many negative as there are positive examples. Once the classifiers for all combinations of path length and path position have been trained, we inspect the resulting feature weights to infer how important each feature is in humans' click choice at each position.

Before presenting the results, we add some notes about the two features. When regression is used for the purpose of feature analysis, it is important to have uncorrelated features. The natural choice for similarity would be the TF-IDF cosine that we have also used in previous sections of this paper. However, this similarity measure is highly correlated with degree: the higher a node's degree, the higher its average TF-IDF similarity with all other articles. This happens because high-degree articles are typically long, and long articles are more likely to have some text overlap with the target article. (The effect is noticeable even in the face of the length-normalization implicit in cosine similarity.) On the contrary, no such correlation with degree is exhibited by the category tree distance. We therefore adopt the latter to quantify similarity in our regression analysis. To be able to compare the weights for features that can take on very different values, we also have to normalize. We do so by adopting a rank-based approach. Consider an article u and a given feature. The neighbors of u get values from the inter-

---

[2]We use complete paths including back-clicks. However, while back-clicks themselves are neglected, we do consider the forward clicks that the player undoes later on.
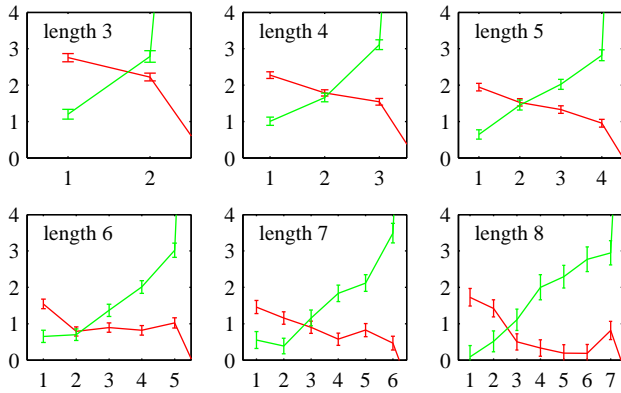
Figure 8: Logistic regression weights for classifying human vs. non-human clicks (with standard errors). *Green:* textual similarity. *Red:* degree. There is one plot per human path length; the *x*-axes show path positions, the *y*-axes weights.

Figure 9: **Overhead with respect to optimal solutions, for single-category *(red)* and most popular multi-category *(blue)* strategies, with one group per target category. The *green* bars show means over all games of the respective target category. From left to right:** PEOPLE, MUSIC, IT, LANGUAGE AND LITERATURE, HISTORY, SCIENCE, RELIGION, DESIGN AND TECHNOLOGY, CITIZENSHIP, ART, BUSINESS STUDIES, MATHEMATICS, EVERYDAY LIFE, GEOGRAPHY.

val $[0, 1]$, such that the highest-ranking neighbor, according to the feature, gets value 1, and the lowest-ranking neighbor value 0.

Fig. 8 plots the resulting weights for the two features. There is one plot for each game length between 3 and 8. The *x*-axes show path positions, and each data point represents one feature weight. The red curves are the degree and the green ones the similarity weights. The weight of the bias term was omitted from the plots, since it is not informative. Note that, for visibility's sake, we do not show the weights for the last click. There, similarity becomes a nearly perfect indicator for the target article, since the target has maximum similarity with itself, so the similarity feature gets a very large weight, and the interesting part of the plots would get squished and hard to read.

Interpreting the plots, our expectation is confirmed. Both features obtain positive weights everywhere, which means that both high degree and high similarity with the target are characteristics of the click choices made by humans. More interesting, as hypothesized, degree dominates in the beginning of games, but as games progress, similarity becomes ever more important, superseding degree starting with the second or third click. Furthermore, similarity starts dominating earlier in more efficient games.

We emphasize that the purpose of this experiment is an analysis of the fitted feature weights, not maximizing the accuracy of the classifiers. Still, to justify our conclusions, we need to show that the classifiers perform better than chance (50%) on a statistically significant level. Evaluating the classifiers on the held-out test set, we find that this is the case. Accuracy is similar for all game lengths. It drops from around 90% for the first path position to about 65% for the second and then stays in the regime of between 55% and 65%. When maximum accuracy is the goal, more powerful features, such as TF-IDF cosine, perform better, but as mentioned earlier, feature correlation does not permit us to use this feature in our analysis.

4.3 Endgame strategies

The main finding of the previous section is that in the opening of games it is common to navigate through hubs. Next we take a closer look at the strategies players adopt in endgames, in order to home in on the target.

In the present analysis, we define an *endgame* as the last 3 articles (i.e., 2 clicks) of a path. To make sure the endgames we analyze do not contain artifacts from the game openings, we consider only games of a full length of at least 5 articles (i.e., 4 clicks). We also neglect all games above the length threshold of 20 articles. The
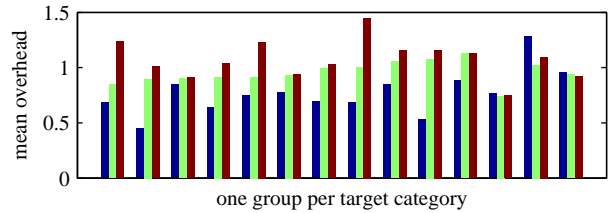
endgame strategy corresponding to an endgame $\langle u_{n-2}, u_{n-1}, u_n \rangle$ is defined as $\langle C(u_{n-2}), C(u_{n-1}), C(u_n) \rangle$, where $C(u)$ is u's top-level category in the hierarchy that comes with our Wikipedia version. For instance, the full category of DIK-DIK is SCIENCE/BIOLOGY/ MAMMALS, and $C(\text{DIK-DIK}) = \text{SCIENCE}$. All 14 values C can take on are listed in the caption of Fig. 9.

We divide the set of all Wikispeedia games into subsets according to target categories, such that all games with target articles from the same category are placed in the same subset. For each target category, we observe between 29 and 104 distinct strategies, out of the possible $14^2 = 196$. For all target categories, the distribution over strategies is highly non-uniform, with most games following one of only a few top strategies. As a consequence, for each target category, the top 10 strategies typically cover between 60% and 90% of all games. Furthermore, the distributions over articles within each category are also non-uniform; e.g., 14% of all instances of GEOGRAPHY are UNITED STATES and 6.1% UNITED KINGDOM.

In 12 out of the 14 target categories, the most popular strategy is the one that consists of the target category only, which we call the 'simple' strategy: people tend to approach the target through articles from the same category as the target. In the remaining two categories, the simple strategy has very high rank, too: it is second most frequent when the target is from DESIGN AND TECHNOL-OGY, and fourth most frequent when it is from PEOPLE. In the former case, the more frequent strategy is $\langle$GEOGRAPHY, GEOG-RAPHY, DESIGN AND TECHNOLOGY$\rangle$; in the latter case, the three more frequent strategies are $\langle$GEOGRAPHY, GEOGRAPHY, PEO-PLE$\rangle$, $\langle$GEOGRAPHY, CITIZENSHIP, PEOPLE$\rangle$, and $\langle$GEOGRAPHY, HISTORY, PEOPLE$\rangle$.

In these examples, GEOGRAPHY seems to play a prominent role. And indeed this is a general property of human paths. To demonstrate this, we count, for each category c, how often articles from it appear in endgames in which the target is *not* also of category c. We find that GEOGRAPHY accounts for 20% of articles in endgames of which GEOGRAPHY is not the target. The next most common categories according to this metric are SCIENCE (7.5%) and HIS-TORY (5.1%). One might argue that certain categories are *a priori* more likely, since they contain more articles. We can correct for this bias by considering the ratio of the above-introduced frequency and the *a-priori* category frequency, i.e., the number of articles in the category, divided by the overall number of articles. In the resulting ranking, GEOGRAPHY is still top, now followed by CITI-ZENSHIP (mostly about politics and culture) and RELIGION. This finding might imply that humans organize their knowledge strongly

according to geographical lines, and that they often associate concepts with their countries of origin.

Previously, we saw that simple single-category endgames are typically most popular with players. Next, we investigate how efficient they are compared to other, more complex strategies. Let l be the length (number of clicks) of a human game, and l^* the number of clicks in an optimal solution. We define the *overhead* of a game as $(l - l^*)/l^*$, i.e., the percentage of the optimal solution length that the information seeker needed extra. For each endgame strategy, we compute the mean overhead over all games of that strategy. As a baseline, we consider the mean overhead across all games of the given target category. We find that the overhead of the simple strategy is on average (over the 14 target categories) 12% higher than that of the mean game; i.e., games using the simple strategy are typically worse than average. Now consider, instead of the simple strategy, the most frequent multi-category strategy. Averaged across all target categories, its overhead is 18% smaller than that of the mean game; i.e., games using the most frequent multi-category strategy are typically considerably better than average. Fig. 9 shows that this is not only true on average but for nearly every target category taken by itself. One of only three categories for which the simple strategy is, on the contrary, most efficient is GEOGRAPHY (the rightmost group in the bar chart). This is in tune with our previous findings: since GEOGRAPHY plays a prominent role even when it is not the target, it makes sense that it allows for efficient paths when it is.

Our interpretation of these findings is that information seekers face a trade-off between efficiency and simplicity. Whilst reaching the target through very related articles from the same category is conceptually simple, the steps taken this way can be small and prolong the game. It often pays off to think out of the box—or to think in geographic terms.

5. TARGET PREDICTION

In the previous sections, we have conducted an in-depth analysis of how people navigate Wikipedia towards a given target article. Our next goal is to apply the lessons learned, in order to design a learning algorithm for predicting an information seeker's target, given only a prefix of a few clicks. Our method explicitly takes the characteristic features of human search into account and is trained on real human trajectories.

There are many potential use cases of such a method. For instance, an intelligent browsing interface could use the algorithm for tracking the user's goal and adapt accordingly, e.g., by suggesting useful shortcuts, thus making human search more efficient. Given the scope of this paper, we evaluate our method only in the context of Wikipedia, but we believe it is general enough to extend to other search scenarios, if appropriate features are used.

Human Markov model. We cast our task as a ranking problem. Given the observed path prefix q, rank all articles t according to how plausible they are as targets of the current search. At the heart of our approach is a Markov model of human search, the parameters Θ of which are learned. To make the prediction, we order candidate targets t according to a ranking function $g(t|q;\Theta)$, defined as the likelihood of t given the prefix q, i.e., as $P(q|t;\Theta)$.

Let $q = \langle u_1,...,u_k \rangle$ be a prefix of $k-1$ clicks. Given target t and model parameters Θ, the probability of seeing q is obtained by multiplying the local click probabilities, and we aim to find the most likely target, i.e.,

$$\arg\max_t P(q|t;\Theta) = \arg\max_t P(u_1) \prod_{i=1}^{k-1} P(u_{i+1}|u_i,t;\Theta), \quad (1)$$

where $P(u_1) = 1/N$ is constant, with N the number of articles (since start articles are picked randomly). Note that we will work with the *prefix log-likelihood* $L(t|q;\Theta) := \log P(q|t;\Theta)$ instead.

In our analysis of humans, we saw that people trade off features differently at different steps. We mimic this by learning a separate set of weights for each step, such that $\Theta = (\theta_1,...,\theta_{k-1})$ is in fact a collection of weight vectors, with θ_i being the weights for step i.

We next propose and test two alternative models of click probability $P(u_{i+1}|u_i,t;\Theta)$, each with its own model fitting algorithm.

Binomial logistic model. The first, simpler model is similar in spirit to the regression of Section 4.2 (but using stronger features), where we fit a model to predict whether humans would pick a given link. The model specifies, for any given click triple (u_i, u_{i+1}, t) separately, the probability that a human would choose it.[3] Formally, we define the *binomial logistic model* as

$$P(u_{i+1}|u_i,t;\Theta) = \frac{\sigma(\theta_i^\top \mathbf{f}(u_i, u_{i+1}, t))}{\sum_{v \in \Gamma(u_i)} \sigma(\theta_i^\top \mathbf{f}(u_i, v, t))}, \quad (2)$$

where $\sigma(x) = (1 + e^{-x})^{-1}$; $\mathbf{f}(u_i, u_{i+1}, t)$ is a feature vector for the click from u_i to u_{i+1} given target t; and $\Gamma(u_i)$ the set of u_i's neighbors. The model parameters Θ are fitted as in standard logistic regression using gradient descent.

Learning-to-rank model. Since the task is to rank target candidates, we also explore a different setup in which we fit Θ explicitly to optimize a ranking objective we refer to as *cumulative reciprocal rank*. This metric is defined as $\ell(r) := \sum_{j=1}^r 1/j$, where r is the rank of the true target [25]. Minimizing this objective implies ranking the true target as high as possible; additionally, cumulative reciprocal rank has the desirable property of putting more emphasis on the top of the ranking (e.g., $\ell(20) - \ell(1) \gg \ell(120) - \ell(101)$), and is therefore a sensible choice for evaluating rankings.

Notice that, unlike assumed by the simplistic binomial logistic model, humans really face a multinomial choice at each step. Therefore, we now represent click probabilities in a *multinomial logistic model*:

$$P(u_{i+1}|u_i,t;\Theta) = \frac{\exp(\theta_i^\top \mathbf{f}(u_i, u_{i+1}, t))}{\sum_{v \in \Gamma(u_i)} \exp(\theta_i^\top \mathbf{f}(u_i, v, t))}. \quad (3)$$

Another advantage of this framework is that we may use other features in addition to the likelihood. Some important factors depend on the entire prefix and cannot be encoded naturally into the Markov model (e.g., How often did the player not take a direct link to the target although this was possible?), which considers only local clicks. Thus, in the learning-to-rank setup, our final ranking function g consists of a linear combination of prefix log-likelihood and those additional *prefix-global features*. As our prediction, we select the target t that maximizes g, i.e.,

$$\arg\max_t g(t|q;\Theta,\beta) = \arg\max_t \beta_1 F_1(q,t) + ... + \beta_m F_m(q,t), \quad (4)$$

where $F_1 = L(t|q;\Theta)$, and $F_2,...,F_m$ are the prefix-global features, the details of which are provided below (note that they do not depend on Θ).

We fit Θ and β using an approach inspired by a method proposed recently by Weston *et al.* [25]. The algorithm minimizes cumulative reciprocal rank via stochastic gradient descent, using a novel sampling trick to speed up learning. In our case, this is necessary since we would otherwise have to iterate over all target candidates

---

[3]One might be tempted to phrase the problem as multinomial logistic regression instead, but this is not possible, since the degree of u_i and hence the number of classes is variable.

(i.e., all articles) for every training example. We refer the reader to Weston *et al.*'s paper regarding the details of the framework and restrict ourselves to highlighting how we adapt their algorithm:

The learning algorithm requires computing the derivative of g with respect to Θ and $\boldsymbol{\beta}$. While $\frac{\partial}{\partial\beta_j}g(t|q;\Theta,\boldsymbol{\beta}) = F_j(q,t)$ is obvious, $\frac{\partial}{\partial\Theta}g(t|q;\Theta,\boldsymbol{\beta})$ is trickier. We forgo a derivation and simply state that

$$
\begin{aligned}
\frac{\partial}{\partial\boldsymbol{\theta}_i}g(t|q;\Theta,\boldsymbol{\beta}) &= \beta_1\frac{\partial}{\partial\boldsymbol{\theta}_i}L(t|q;\Theta) \\
&= \beta_1\Big[\mathbf{f}(u_i,u_{i+1},t)-\sum_{v\in\Gamma(u_i)}P(v|u_i,t;\Theta)\,\mathbf{f}(u_i,v,t)\Big].
\end{aligned}
$$

That is, the likelihood gradient with respect to the weights for prefix position i is equal to the difference of the feature vector of click i and the expected feature vector under the current weights Θ.

Features for learning. So far, we have only described the abstract framework of our algorithms but have not yet discussed the concrete features we use. As mentioned, our choice of features is inspired by the results of our study of human behavior: if we design features that capture the characteristics of human paths, the algorithms can learn weights to predict targets under which the seen prefix resembles human behavior most. Specifically, these are the entries of the likelihood feature vector $\mathbf{f}(u_i, u_{i+1}, t)$ (recall that we learn a separate weight vector for each i):

1. TF-IDF(u_{i+1},t): as we have seen, articles become ever more related to the target as human games proceed (Fig. 6 (3,2));
2. TF-IDF(u_i,u_{i+1}): 'textual step size' is large at first and becomes ever smaller (Fig. 6 (3,3));
3. deg(u_{i+1}): humans commonly navigate through hubs in the beginning, so we expect the weight for this feature to be large for early and small for later steps i;
4. deg$(u_i) \times$ deg(u_{i+1}): if u_i already is a hub, there is no more need for the player to search for another one, which can be captured by this interaction feature;
5. TF-IDF$(u_i,t) \times$ deg(u_{i+1}): if the player is already close to t textually, finding a hub is less important;
6. TF-IDF$(u_i,t) \times$ TF-IDF(u_i,u_{i+1}): close to t, the step size is typically also smaller;
7. the indicator SPL$(u_{i+1},t) >$ SPL(u_i,t): if a click increases the shortest-path length to t, then it is less likely to be chosen by a player;
8. $|$TF-IDF$(u_i,t) -$ TF-IDF$(u_{i+1},t)|_+$: a click is also less likely if it decreases the textual similarity to t.

In the learning-to-rank approach, we additionally have the following prefix-global features (cf. (4)):

1. Number of times the player could have taken a direct link to t yet did not;
2. Number of clicks through which the SPL to t increased;
3. Sum (over all clicks) of decreases in textual similarity to t.

The last two prefix-global features are similar to likelihood features 7 and 8, but here they can modify the ranking function explicitly rather than merely via the likelihood term. We expected the first prefix-global feature to receive a large negative weight, guided by the intuition that humans would always go directly to the target as soon as this is possible. However, a weight of nearly zero is learned for this feature, which indicates that information seekers often miss the best links because they do not expect them and hence do not notice them in the often long article text. This emphasizes the usefulness of the task of target prediction: if the algorithm can
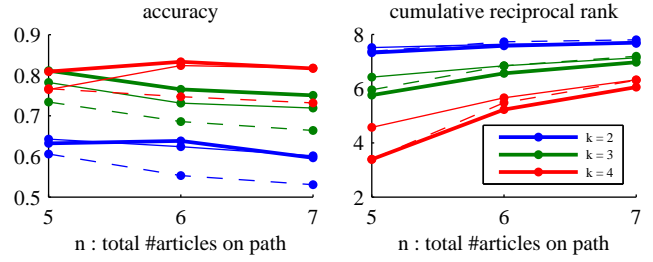

Figure 10: Performance of our target prediction algorithms, for varying prefix lengths k (indicated by color). *Left:* accuracy (higher is better). *Right:* cumulative reciprocal rank (lower is better). *Bold solid:* multinomial ranking model. *Thin solid:* binomial logistic regression. *Dashed:* TF-IDF baseline.
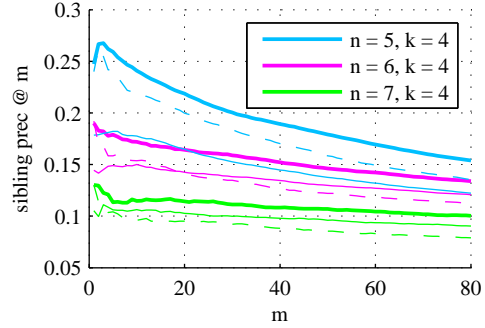

Figure 11: Sibling precision of our target prediction algorithms. *Bold solid:* multinomial ranking model. *Thin solid:* binomial logistic regression. *Dashed:* TF-IDF baseline.

infer the intended target, it could highlight the most useful links so they become more salient to the user.

Evaluation. For training our models, we use prefixes of 3 clicks, taken from human games of at least 4 clicks. The weights converge quickly for the multinomial ranking model, after seeing a few thousand examples, which takes only some minutes.

We also compare our algorithms to a baseline that simply predicts as the target the article with the largest TF-IDF similarity to the last article of the prefix.

We use a test set not seen during training to evaluate the algorithms on two tasks: (1) given a prefix $q = \langle u_1,...,u_k\rangle$ and a choice of two targets, pick the true target t; the false target is picked randomly from the set of articles that ever occurred as targets and have the same SPL from u_k as t; (2) given q, rank the set of *all* articles such that the true target is ranked high. The second task is much harder but also more useful than the first; e.g., if the method is to be implemented for an intelligent browsing tool, reasonable targets must be picked from all candidates, not just from a set of two.

In the first task, we use accuracy as a metric; in the second, we measure ranking loss according to cumulative reciprocal rank, the objective we also use for training. While this captures ranking quality objectively, it might be overly strict; e.g., if $t =$ WINE, then predicting BEER is much better than, say, GASOLINE. We account for this by measuring 'sibling precision@m', which is the same as precision@m, with the difference that not only t but all articles from the same category as t are counted as relevant (we use the leaves of the hierarchy of our Wikipedia version as categories).

We vary two parameters of the test prefixes: k, the number of articles in the prefix; and n, the length of the entire human path. The results are summarized in Fig. 10 and 11. In all plots, the bold solid

line represents the multinomial ranking (MR) model, the thin solid line the binomial logistic regression (BLR) model, and the dashed line the TF-IDF baseline. First note that MR is at least as good as, and often better than, both BLR and TF-IDF according to every metric. Now consider Fig. 10. As expected, our methods work better when prefixes are longer (cf. the order of the bold curves) and when full paths are shorter (cf. the slopes of the curves). Notably, on the task of picking the correct one of two targets, MR achieves an accuracy of 80% when 3 clicks are seen, regardless of whether the entire game is 4, 5, or 6 clicks long. Interestingly, while BLR has higher accuracy on the binary task, the simple TF-IDF baseline achieves better ranking performance. We take this as an indicator that MR combines the better properties of both.

Finally, consider Fig. 11, which shows sibling precision@m. For the sake of brevity, we display only the case $k = 4$, but in relative terms the results are the same for all prefix lengths. The precision@30 of MR is 20% for $n = 5$, which means that 6 of the top 30 targets are of the same category as the true target, when we see 3 clicks and the full game has 1 more click. Even when there are 2 (3) more clicks, we still see 5 (3) top-ranked articles that are very close to the true target (for comparison, in a random ranking, precision is only 1% on average). This property of the ranking algorithm is desirable, since in a real-world application making a close enough guess might often be nearly as good as predicting the exact target.

6. CONCLUSIONS

Finding paths connecting different concepts—like linking causes to effects—is a task the human race has been performing for millennia. We formalize this task in a human-computation game of wayfinding between the concepts of Wikipedia. We study more than 30,000 goal-directed human search paths and identify aggregate strategies people use when navigating information spaces. As information spaces become more complex, it is increasingly important to understand how humans navigate them and to assist them in locating the desired information. This is the second focus of our paper, where we build a predictive model of human wayfinding that can be applied towards intelligent browsing interfaces.

The view of human wayfinding as a navigation task on Wikipedia points to a broad range of interesting issues, and our goal in this paper has been to start exploring the foundations for reasoning about these questions. We anticipate further investigations in determining why people give up navigating and characterizing unfinished wayfinding tasks. Given the insights we offer here, another interesting direction is in automatically designing information spaces that humans find intuitive to navigate and identifying individual links which could make the Wikipedia network easier to navigate. Overall, we hope that this perspective can contribute to the development of new functionality in the continuing evolution of how we use and navigate the Web.

7. ACKNOWLEDGEMENTS

8. REFERENCES

[1] L. Adamic and E. Adar. How to search a social network. *Social Networks*, 27(3):187–203, 2005.

[2] E. Adar, J. Teevan, and S. T. Dumais. Large scale analysis of Web revisitation patterns. In *CHI*, 2008.

[3] M. J. Bates. The design of browsing and berrypicking techniques for the online search interface. *Online Review*, 13(5):407–424, 1989.

[4] M. Bilenko and R. W. White. Mining the search trails of surfing crowds: Identifying relevant websites from user activity. In *WWW*, 2008.

[5] E. H. Chi, P. Pirolli, K. Chen, and J. Pitkow. Using information scent to model user information needs and actions and the Web. In *CHI*, 2001.

[6] P. S. Dodds, R. Muhamad, and D. J. Watts. An experimental study of search in global social networks. *Science*, 301(5634):827–829, 2003.

[7] D. Downey, S. T. Dumais, and E. Horvitz. Models of searching and browsing: Languages, studies, and applications. In *IJCAI*, 2007.

[8] D. Helic, M. Strohmaier, C. Trattner, M. Muhr, and K. Lerman. Pragmatic evaluation of folksonomies. In *WWW*, 2011.

[9] P. Killworth, C. McCarty, H. Bernard, and M. House. The accuracy of small world chains in social networks. *Social Networks*, 28(1):85–96, 2006.

[10] J. M. Kleinberg. Navigation in a small world. *Nature*, 406(6798):845–845, 2000.

[11] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *PNAS*, 102(33):11623–11628, 2005.

[12] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[13] S. Milgram. The small-world problem. *Psychology Today*, 2(1):60–67, 1967.

[14] J. Muramatsu and W. Pratt. Transparent Queries: Investigating users' mental models of search engines. In *SIGIR*, 2001.

[15] V. L. O'Day and R. Jeffries. Orienteering in an information landscape: How information seekers get from here to there. In *CHI*, 1993.

[16] C. Olston and E. H. Chi. ScentTrails: Integrating browsing and searching on the Web. *TCHI*, 10(3):177–197, 2003.

[17] P. Pirolli and S. K. Card. Information foraging. *Psychological Review*, 106(4):643–675, 1999.

[18] A. J. Sellen, R. Murphy, and K. L. Shaw. How knowledge workers use the Web. In *CHI*, 2002.

[19] Ö. Şimşek and D. Jensen. Navigating networks by using homophily and degree. *PNAS*, 105(35):12758–12762, 2008.

[20] A. Singla, R. W. White, and J. Huang. Studying trailfinding algorithms for enhanced web search. In *SIGIR*, 2010.

[21] 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 *CHI*, 2004.

[22] R. H. Trigg. Guided tours and tabletops: Tools for communicating in a hypertext environment. *TOIS*, 6(4):398–414, 1988.

[23] R. West. Wikispeedia. Website, 2009. http://www.wikispeedia.net (accessed Feb. 2012).

[24] R. West, J. Pineau, and D. Precup. Wikispeedia: An online game for inferring semantic distances between concepts. In *IJCAI*, 2009.

[25] J. Weston, S. Bengio, and N. Usunier. Large scale image annotation: Learning to rank with joint word-image embeddings. *Machine Learning*, 81(1):21–35, 2010.

[26] R. W. White and J. Huang. Assessing the scenic route: Measuring the value of search trails in Web logs. In *SIGIR*, 2010.

[27] Wikipedia. 2007 Wikipedia Selection for schools. Website, 2007. http://schools-wikipedia.org (accessed Aug. 2008).