Remodeling Information Networks based on Collective Human Intelligence

Ashwin Paranjape
Computer Science Department
Stanford University
ashwinp@cs.stanford.edu

1. INTRODUCTION

Human knowledge is often represented as an information network, be it traditional encyclopedias with references to related terms or journal publications with citations, or be it simply the whole internet with links to related (or perhaps unrelated) concepts.

Before the advent of Wikipedia, information networks were curated by a small set of subject experts. Thus the links between information nodes were based on their individual understanding of the network. Wikipedia, which is a free-access and free-content Internet encyclopedia, however allows anyone to edit almost all of its articles and thus depends heavily on community participation for creation and maintenance. This unique quality allows its English version, according to its Wikipedia page, to contain thirty times as many articles as Britannica Encyclopedia and still maintain a level of accuracy which approaches it.

Information nodes in larger networks like Wikipedia are even harder to connect because, this task is the collective effort of individuals from the community and each person has local knowledge of the information network. Also the connection is based upon his or her own individual understanding, which may not be the same as others, and more importantly, may not be useful for the community.

There are two subproblems which fall in the domain of introducing new links in Wikipedia link graph.

Target prediction: The task here to identify possible target pages to which a source page should link. For example, consider that an editor has created a new Wikipedia article and wants to link to other existing articles to make the new article more understandable. The task is to identify possible phrases in his text which need further elaboration and find articles which do so.

Source prediction: In contrast, the task here to identify possible source pages which should link to this target page. For example, consider that an editor has created a new Wikipedia article. However, it is of little use to other articles (sources) which contain phrases which can be explained by this new article (target), until such links are created in the sources. However the task of determining the usefulness of the new article for each source which mentions it is time consuming and sought with errors.

Due to the inherenet directionality of hyperlinks the source prediction problem is harder because the candidate set of sources is very large compared to the single page which needs to be looked at for target prediction.

Our task here is to automate the source prediction task. There are many methods which introduce new links based on either the link graph or based on textual or categorical understanding of content. However what is missing from the picture is the realization that there are additional sources of data which can be harnessed. One such data source is human navigational traces.

Present Work: Navigation logs for source prediction. Human navigational traces are captured by usage logs recorded on the server side by many websites. These logs contain strong signals with regard to which existent hyperlinks are better suited for human navigation. Extending this notion further, they would also contain clues to whether a non-existent link should be introduced. If we often observe users going through pages s and ending up in page t then although s does not link to t, we should probably introduce a new link or a shortcut.

Present Work: Proposed approach. We propose to use Wikipedia as our proof-of-concept domain because high-quality navigation logs are available for it and Wikipedia has a well maintained link structure which is also continuously evolving. This link history gives us the opportunity to automatically generate ground truth based. The high quality navigation logs come from a class of human-computation games known as Wiki-racing. In these games the user has to navigate from a given source to a target in least number or steps. The key point here is that the underlying link structure is known to humans on a local level. Thus this is an instance of a decentralized search based on human intuition.

Building on the above intuition, if a page s is traversed by many users in search of target t then this is an indicator that users expect s to link to t. If such a link does not exist but s contains a mention of t, then s is a possible candidate.

As a concrete example, consider Fig. 1. The figure summarizes several navigational paths, all with the target $t =$ INFLAMMATION. Paths progress from bottom to top, and only the last few clicks are shown per path. Each node s also contains the fraction of all paths with target INFLAMMATION that passed through s. For instance,

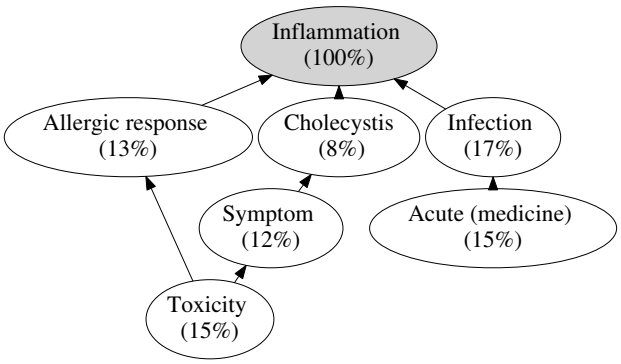

Figure 1: In this graph, the target t is INFLAMMATION. Other nodes are articles s in the Wikipedia graph that appeared on paths to t. The number for each s is the percentage of paths with target t that passed through s.

we see that 17% of times INFLAMMATION was reached from INFECTION and 13% of times it was reached from ALLERGIC RESPONSE. A considerable fraction of paths (15%) passed through ACUTE (MEDICINE), which does not link to t, although it mentions t several times and could clearly benefit from a link to it.

The central part of our approach is that we mine many link candidates (s,t) from a large number of navigation traces for each target t and then rank these candidates by relevance.

2. RELATED WORK

This section has been kept short because we can only summarize in the limited space and also because it was covered covered in detail in the reaction paper.

There is a rich line of work on identification of missing links among Wikipedia articles or linking an existing webpage to Wikipedia [5, 6, 7, 18]. Generally, these approaches focus on building models of the Wikipedia graph structure while also performing keyword extraction and word sense disambiguation [9, 10, 15].

Some very relevant work was done by Ageev et al. [2]. They, too, have developed a human-computation game for collecting data, in which users are asked to find the answers to as many factual questions (e.g., 'What is the highest peak in the Western Hemisphere?') as possible within a given amount of time, using Web search queries that may optionally be followed by click-based navigation. There, too, the goal is explicitly known, but not in the form of a specific target page but rather in the form of a specific answer string. Ageev et al.'s setup has the advantages of a more realistic search scenario, since both querying and navigating are allowed, and of a more realistic browsing environment (the Web rather than Wikipedia only). On the flip side, our approach allows for several types of analysis that are impossible in their setup. For instance, each Wikipedia page has a clearly defined topic, with the target being exactly one page (rather than an answer string that might be gleaned from several pages), such that it is straightforward to quantify how related a page along the click path is to the target. Moreover, the network topology is known entirely in our case.

3. DATASETS

The data set of Wikipedia navigation traces was collected through a popular online game that is generically known as 'Wikiracing' [17]. Several websites offer versions of this game, such as *Wikispeedia* [14, 12] or *The Wiki Game* [3], but they all share the same general idea: a user is given two Wikipedia articles—a *start* and a *target*—and is asked to navigate from the start to the target by exclusively clicking hyperlinks contained in the visited pages. We also refer to start–target pairs as *missions*.

3.1 Wikispeedia[11]

This consists of around 32,000 paths which were collected from human navigation on Wikipedia articles. Each user was given a mission consisting of a source and a target article, sampled randomly from the set of all articles in the strongly connected component of Wikipedia's hyperlink graph. Alternatively the user can choose a mission from a list of 5 randomly sampled missions solved previously by other users. The task is to navigate in as few clicks as possible by following hyperlinks. Backtracking using back button in browser is allowed and the user is also allowed to view the target article in case he doesn't know about it.

A static, condensed and curated version of Wikipedia [1] containing 4,600 articles and around 120,000 links is used. Another important thing to note is that this version contains no redirects, making data processing much simpler. The data is collected since August 2009, with voluntary participation from Web users. Also only 49% of these paths completed successfully.

3.2 The Wiki Game[4]

The wiki game is a website which contains a game called speed-racer, which generates a new mission every 120 seconds, from a randomly selected source and target Wikipedia article. All the users playing speed-racer, race against each other to navigate from source to target using hyperlinks on Wikipedia. Another allied game is '5 clicks to Jesus' where the target is fixed.

The advantage of using this dataset is that it has a much larger size than Wikispeedia. After correcting for redirects and vandalism, it contains 1.1 million paths over 460,000 missions, which should give us much stronger signals about the missing links.

We work with a set of 974k paths collected from 2009 through 2012. For each path, the data set also contains the user id, the type of game (cf. above), and timestamps for all clicks. Paths group into 364k distinct missions (start–target pairs), i.e., there are 2.7 paths per mission on average. The number of distinct targets is 3k, i.e., we have 325 paths per target on average, with a median of 208. Fig. 2 shows a complementary cumulative distribution function of the number of paths per target (excluding the target JESUS, for which there are 67k paths). We note that targets with many paths are quite frequent. In the following, we focus on the 2,002 targets (67% of all targets) that have at least 100 paths.

4. METHOD

The earlier plan was to look for strong signals such as backtracking, but after some data analysis it seems that such signals are rare and not so useful.

We now take an approach which is minimalistic in its assumptions about the browsing patterns and is robust because a large amount of data is applicable to this approach. The hypothesis is that if a page s is traversed by many users in search of a target t, then this is

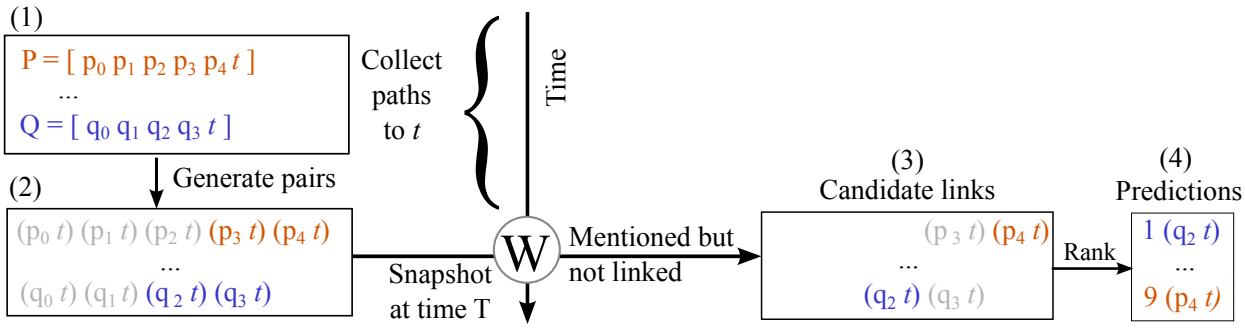

Figure 3: Overview of our approach. (1) Collect a set of paths to a target t till time T and capture the current Wikipedia snapshot at T. (2) Create (s,t) pairs, excluding the first half of all paths. (3) Filter those (s,t) pairs based on the current Wikipedia snapshot to generate as candidate links those (s,t) pairs that do not correspond to a link in the snapshot. (4) Rank the candidate links.

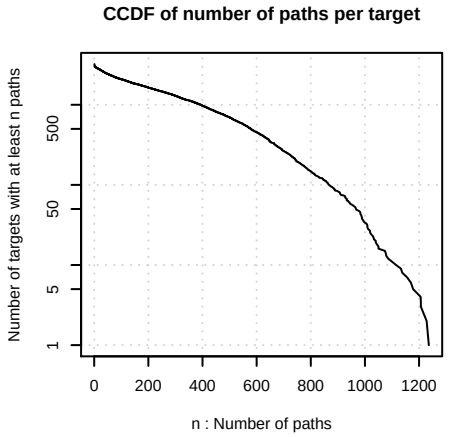

Figure 2: Complementary cumulative distribution function of the number of paths per target in the navigation trace data set from The Wiki Game, on a logarithmic y-axis. We use only those targets which have at least 100 paths.

an indicator that humans expect the link from s to t to exist. So if the link between s and t does not exist then but s contains a phrase that can be used as an anchor for t then we should consider it as a candidate.

So we have a filtering mechanism in place to select candidates. In contrast, the null hypothesis does not select any candidates.

An outline of our method for source prediction is given by Fig. 3. We start by collecting navigation traces up to the current time T. For each path $p = \langle p_0, \ldots, p_n = t \rangle$, the initial set of candidates is $\{(p_i, t) \,:\, 0 < i < n\}$, i.e., every direct link from any page along the path to the target t is initially a candidate (the start page p_0 is exempt, since it is selected randomly and is therefore unlikely to be a useful candidate).

There are in general many paths for the same target t (upper left box in Fig. 3), so we take the union of the candidate sets resulting from all these paths as the initial candidate set for t (lower left box in Fig. 3).

We consider the Wikipedia snapshot $\mathcal{W}$ at the current time T. A link (s,t), where $s \in \{p_1, \ldots, p_{n-1}\}$, can be suggested only if it does not already exist in $\mathcal{W}$. Further, the source s should mention a phrase that could serve as the anchor for a link to t; that is, s should mention t.[1]

To detect pages that mention the target t we construct the set $\mathcal{M}_t$ of all phrases that serve as anchor texts for t across all articles in the current Wikipedia snapshot.[2] We then say that s mentions t if it contains any phrase from $\mathcal{M}_t$. As has been highlighted in brainstorming section, there are many signals for link prediction.

4.1 Source candidate ranking

Source candidate selection yields an unordered set of candidates for each target t. The goal of the next (and final) step in our pipeline is to turn this set into a meaningful ranking. Since the source prediction task (Fig. 1) asks for sources for a given target t, we produce a separate ranking for each t. Several ranking are conceivable:

1. **Ranking by relatedness.** It seems reasonable to rank source candidates s by their relatedness to t, since clearly a link is more relevant between articles with topical connections.[3] Relatedness may be measured in various ways, e.g., by the number of mentions s makes of t or by the TF-IDF cosine of s and t, to name but a few. Some of the best relatedness measures are based on Wikipedia, and since we deal with Wikipedia as our data set, we choose this option.

2. **Ranking by frequency.** Navigation traces provide us with statistics about how frequently a source s was traversed by users searching for target t. Based on this, we compute the *path frequency,* the fraction, out of all paths with target t, of paths that also passed through s.

[1] But cf. Sec. **??** for a discussion of how a suggestion could be useful even in the absence of a mention.

[2] In practice, we exclude (1) phrases that rarely (less than 6.5% of all cases [9]) serve as link anchors for any target, which excludes, e.g., 'A' as an anchor for AMPERE, and (2) anchor texts for which t is seldom (less than 1% of all cases) the target, which excludes, e.g., 'Florence' as an anchor for FLORENCE, ALABAMA.

[3] According to the Wikipedia linking guidelines [16], links should correspond to 'relevant connections to the subject of another article that will help readers understand the article more fully.'

We experiment with two relatedness measures for case 1 above. The first is due to Milne & Witten [8] and is based on the inlink sets $\mathcal{S}$ and $\mathcal{T}$ of s and t, respectively. It calculates the relatedness of s and t as the log probability of seeing a link from $\mathcal{S} \cap \mathcal{T}$ when randomly sampling a link from the larger one of the sets $\mathcal{S}$ and $\mathcal{T}$ (normalized to lie between 0 and 1):

$$\mathrm{MW}(s,t) = \frac{\log(|\mathcal{S} \cap \mathcal{T}|) - \log(\max\{|\mathcal{S}|, |\mathcal{T}|\})}{\log(\min\{|\mathcal{S}|, |\mathcal{T}|\}) - \log(N)}, \qquad (1)$$

where N is the total number of Wikipedia articles.

The second relatedness measure is due to West et al. [15] and works by finding a low-rank approximation of Wikipedia's adjacency matrix via the singular-value decomposition (SVD). The pair (s,t) corresponds to an entry $A[s,t]$ in the adjacency matrix A and to an entry $A_k[s,t]$ in the rank-k approximation A_k obtained from A via SVD. If $A[s,t] = 0$ and $A_k[s,t] \gg 0$, then s does not link to t yet but is a good candidate. So we define the SVD-based relatedness as

$$\mathrm{SVD}(s,t) = A_k[s,t] - A[s,t]. \qquad (2)$$

5. EXPLORATORY ANALYSIS OF LINK CANDIDATES

Having introduced the data set and our source prediction method, we now explore the data set of human navigation traces to build intuitions on strengths and potential weaknesses of our approach.

Number of pages on a path mentioning the target. We count for each path $p = \langle p_0, \ldots, p_n = t \rangle$ how often the target t is mentioned across all visited nodes $p_1, \ldots, p_{n-1}$ (excluding the randomly selected start page p_0) and find that, on average, t is mentioned 1.7 times per path. Since p_{n-1} contained a link to t, it is very likely to also mention t (for our definition of a mention, cf. Sec. 4), which means that, on average, each path contains 0.7 additional pages that mention t.

Now consider the subset of visited pages that mention t. Out of these, 73% contain a link to t in the current Wikipedia snapshot. The remaining 27%, which do not link to t in the current Wikipedia snapshot $\mathcal{W}$, are potentially good candidates sources to link to t, since these pages were actively chosen by the user while she was searching for t.

Information along the path. We also investigate which parts of a path carry most value for source prediction. Consider Fig. 4, which aggregates all paths and shows for each part of the path how likely the pages in that part are to mention t. In order to be able to aggregate paths of variable length, we adopt the notion of relative path position: the relative position of p_i along the path $p = \langle p_0, \ldots, p_n = t \rangle$ is i/n. Fig. 4 uniformly buckets the range $[0,1]$ into 5 intervals and plots the average for each interval. We only include paths of at least 5 clicks, such that each path contributes to each bucket, and the page p_{n-1} just before the target always falls into the last bucket.

We see that target mentions become more frequent as paths progress (the black curve in Fig. 4). We are particularly interested in mentions that are not accompanied by a link (the magenta curve in Fig. 4), since these are our source candidates. So the figure tells us that candidates are more likely to appear towards the end of paths.

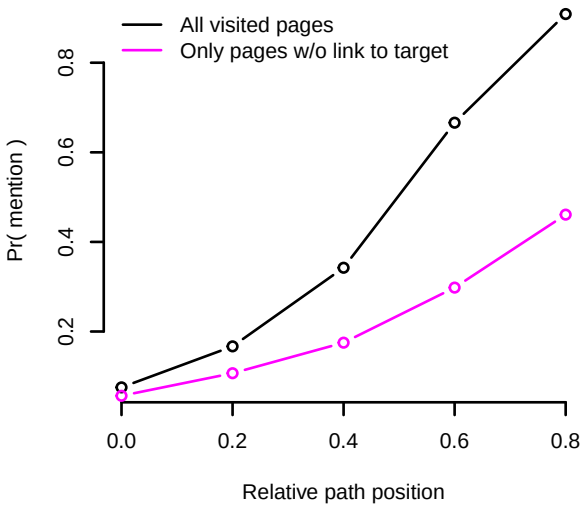

Figure 4: Target mentions become more frequent as paths progress. The magenta curve represents candidate mentions, i.e., mentions that are not accompanied by a link in the current Wikipedia snapshot. We observe that candidates become more frequent as we get closer to target.

We note that these curves are in tune with previous work [13], which has shown that humans tend to follow a 'semantic gradient' during information network navigation, passing through articles that get ever more related to the target. Hence it makes a lot of sense that the rate of target mentions should increase as paths progress.

In order to form intuitions about how meaningful our suggestions are, we would ideally like to evaluate for each relative path position how good the source candidates at that position are. However, ground truth data is hard to come by; in order to make strong claims, we need to ask humans how good our predictions are (Sec. 6). Since this is expensive and time-consuming, we adopt a notion of ground truth that is approximate and biased but nevertheless allows us to gain some initial insights.

5.1 Obtaining ground truth based on Wikipedia evolution

We obtain a weak notion of ground truth from the evolution of Wikipeda graph structure.We label a source candidate s as positive if the link (s,t) existed for at least an α-fraction of the total lifetime of the article s.[4] Source candidates s with a true label correspond to links (s,t) that existed for a substantial amount of time but got deleted before the current Wikipedia snapshot $\mathcal{W}$. That is, such a link would have been valuable for navigation yet was removed at some point in time, and we should consider reintroducing it.

More precisely, we consider some page s to be a valid source suggestion: If the link (s,t) did not exist during the game then the user

[4]We computed these values based on the complete Wikipedia edit history.

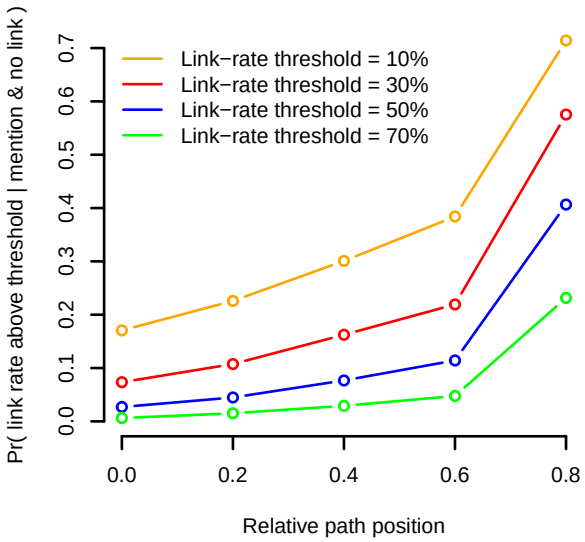

Figure 5: Fraction of positive links from automatically generated ground truth becomes higher for pages appearing later on in paths. This trend holds for several thresholds.

could not have taken it although she might have intended to do so; in this case, (s,t) would clearly be a good link suggestion. If the link did exist during the game but was not clicked by the user, this means that she did not see it in her rush to reach t as fast as possible (or else she would have clicked on it to immediately win the game); so either the user found another promising way to continue the search before seeing the link to t, or the link was too hard to find in the text of s, which in turn would be a signal that we should make that link more obvious. Finally, if the link existed during the game (and was possibly even clicked by the user), but has been deleted since, then it is probably a good idea to suggest it for reintroduction.

Fig. 5 captures this notion of candidate quality, again broken up by relative path position. The graph shows that the fraction of positives obtained from automatically obtained ground truth becomes higher for pages appearing later on in paths. We conclude that not only are mentions at later positions more frequent (Fig. 4), they also correspond to better link anchors. (We try several values for the threshold α but the same trend holds for all thresholds.) This provides additional justification for our decision to only include the second half of navigation traces in our set of source candidates (Sec. 4).

6. EVALUATION

In our experiments, we compare five methods: Given a target t, we can either consider as source candidates the set of all articles that mention t, across all of Wikipedia, but do not link to it; or we can subselect candidate sources based on whether we observe them in navigation paths (Sec. 4). Further, we consider two relatedness measures for ranking (Sec. 4.1). This yields four combinations of candidate selection methods ('none' and 'path-based') and relatedness measures ('MW' and 'SVD'). The fifth method requires no external relatedness measure but simply ranks candidates with respect to their frequency among paths with target t (Sec. 4.1).

To sum up, we consider the following five methods for predicting missing links to a given target page t:

- **All, rank by MW:** We use all candidate sources and rank based on the M&W method.
- **All, rank by SVD:** We use all candidate sources but rank using the SVD method.
- **From paths, rank by MW:** Only use candidates based on navigational traces and then rank them based on the M&W method.
- **From paths, rank by SVD:** Only use candidates based on navigational traces and then rank them based on the SVD method.
- **From paths, rank by frequency:** Only use candidates based on navigational traces and then rank them based on the frequency they appear in paths.

Experimental setup. We perform a twofold evaluation, one based on the automatically obtained and approximate labels defined in the previous section, the other based on labels obtained from human raters.

We refer to our automatically obtained labels as 'weak' because, by definition, they contain many false negatives. Wikipedia is an evolving organism, and an important part of our task is to suggest links which never existed. However, by the α-threshold criterion defined in Sec. 5, these links will be counted as negative examples. For example, the article on ACUTE (MEDICINE) should clearly link to INFLAMMATION, as it explains a concept critical to understanding the term ACUTE as used in medicine, but the link (ACUTE (MEDICINE), INFLAMMATION) is labeled as negative by the automatic ground truth, since ACUTE (MEDICINE) has never linked to INFLAMMATION in Wikipedia's history.

Nonetheless, the automatic ground truth is useful during development because it provides us with many labeled examples for free and allows for relative comparisons between different methods.

We then select a subset of targets, predict sources for them using the methods that performed best during the development phase on the automatic ground truth, and ask humans on Amazon Mechanical Turk to label the top predictions. Here we get rid of the shortcomings of the automatic ground truth, on which we cannot obtain absolute performance numbers (mainly due to the high false-negative rate), but have less data to work with.

Evaluation metrics. As our evaluation metric, we use Precision@k for $k = 1,\ldots,10$. This simplifies the process of generating human-labeled ground truth because it means that we only need to request labels for the top k predictions for each compared method.

As outlined in Sec. 4, a link to t may be suggested for any source s that mentions t in the current Wikipedia snapshot $\mathcal{W}$ but does not have a link to it there. Since we evaluate precision@k for $k = 1,\ldots,10$, we can further only include targets for which our method

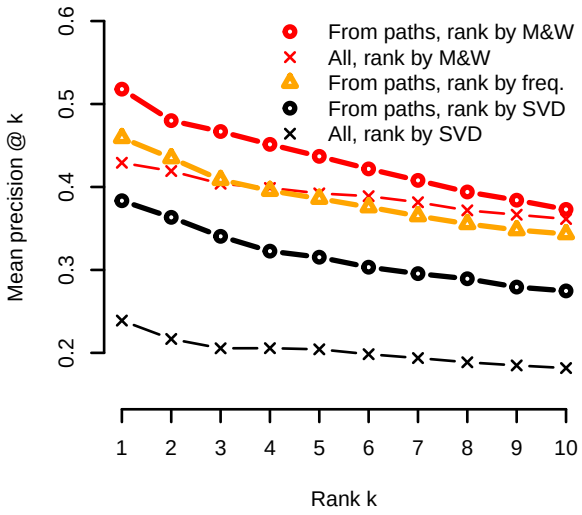

Figure 6: Precision@k for different methods and rankings. Bold lines represent path based candidate selection. Notice that "Paths" consistenly perform better than "All".

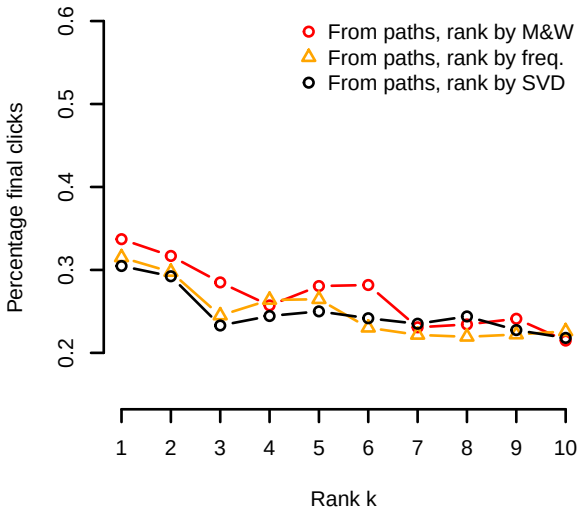

Figure 7: Top suggestions are often reintroducing links which were taken by the user as the final click into the target but are not present in Wikipedia.

finds at least 10 source candidates, which defines our evaluation set of 699 targets.

6.1 Evaluation using automatically obtained ground truth

The precision@k curves for all five evaluated methods are displayed in Fig. 6, and their performance is summarized in terms of the area under the precision@k curve in Table 1.

Overall, we achieve good performance. Especially given that our ground-truth is of high precision but low recall. Even though the Precision@k decays from 0.5 at $k = 1$ to 0.4 at $k = 10$ manual inspection of errors of the algorithm actually revels that the suggested links make sense and are truly missing. (And the Wikipedia community has just not yet found time to include these links into the Wikipedia graph so that they would make their way into our ground truth set).

Comparing different methods we observe that path-based candidate selection performs better than doing no subselection for both relatedness measures used in ranking. Path-based selection improves performance by a particularly large margin for the SVD-based ranking method, which has much lower precision@k than the other methods. This establishes the fact that there is a lot of value in path-based candidate selection especially when the ranking measure by itself is not up to the mark.

The margin between path-based selection and no selection is larger for smaller k, which means that considering navigational paths is particularly useful for predicting the top link sources.

Note that both ranking measures (MW and SVD) make use of the high-quality link structure of the Wikipedia page graph (Sec. 4.1). If we wanted to generalize our approach to domains beyond Wikipedia, we can easily imagine scenarios where no such high-quality ranking measures are readily available (e.g., when pages are not as topically coherent as Wikipedia articles, or pages have scarce content and are poorly interlinked). With such situations in mind, it is encouraging to see that our fifth measure ('ranking by frequency') in combination with path-based candidate selection (the yellow curve in Fig. 6) performs quite competitively.—Recall that that ranking measure does not rely on any external relatedness measure but simply ranks candidates with respect to the frequency with which they appeared among the paths with target t. This is an important observation because it means our method has the potential generalize well to use cases where a good ranking measure is not readily available.

By construction, the last click on a path always leads into the target. The fact that a user looked for, found, and clicked on this link is a very strong signal that the link is useful for navigation. Removing such links from Wikipedia is particularly harmful from a user-interface perspective, and it is desirable that a source prediction method suggest them for reintroduction. To see if our path-based candidate selection method meets this desideratum, Fig. 7 plots, for each k, the fraction of predicted links that were also the last link on the paths they were mined from. We conclude that our top suggestions are often links that were taken by the user as the last link to tar-

Candidate selection	Rank by MW	Rank by SVD	Rank by path freq.
None	39%	20%	N/A
Path-based	43%	32%	39%

Table 1: Area under the Precision@k curve for no candidate selection versus path-based candidate selection for all ranking measures. Note that path frequency is only applicable for path based candidate selection.

get and thus our method rightly reintroduces such links back into Wikipedia.

6.2 Evaluation based on human raters

Motivation for human evaluation. Wikipedia is a continuously evolving entity. Although the link history, on the basis of which we defined our automatic ground truth, captures its evolution, it can only tell us which links are positive examples (because they persisted throughout long period of time). However, this notion of ground truth suffers from false negatives, since there are many links that should be, but have never been, added. So if our method suggests a link that should have been there but was never yet introduced, then the previous evaluation would count it as a bad suggestion. At the end of this section, we confirm the prevalence of false negatives *post hoc,* after having collected ground-truth labels from humans.

Methods compared via human evaluation. In our human evaluation, we compared the two top-performing methods: (1) path-based candidate selection with MW ranking and (2) no candidate selection with MW ranking. By using the same ranking method and only varying whether path-based candidate selection was performed, we can gauge the impact of the latter on predictive performance.

Target sampling. In order to select targets for which to evaluate the predictions of the two methods, we stratify the base set of 699 targets by the number of paths observed per target and select 10 targets from each decile. The rationale behind stratification is that we want to avoid introducing a bias while subselecting by potentially being skewed towards targets for which the path-based candidate selection method made particularly large number of predictions.

Obtaining ratings through Amazon Mechanical Turk. We use Amazon Mechanical Turk for recruiting human raters. As in the automatic evaluation, our goal is to assess the precision@k, where $k = 1, \ldots, 10$, for the two compared methods. In each rating task, the human evaluator was presented with a target t and a set of (up to) 14 candidate sources and was asked to guess which of the candidate source articles were likely to contain a link to target article. There were no constraints on the number of source articles to be chosen. The set of (up to) 14 candidate sources comprised the following entries:

- 5 predictions from each of the two compared methods (either suggestions 1 through 5 or suggestions 6 through 10 from each method),
- 2 control sources, sampled randomly from the set of all Wikipedia articles that link to t,
- 2 control non-sources, sampled randomly from the set of all Wikipedia articles, which made them highly unlikely to link to the target.

In cases where the two methods agreed on a suggestion, that suggestion was included only once in the set of source candidates. Also, to prevent any ordering bias, we shuffled the order sources in the presented list.

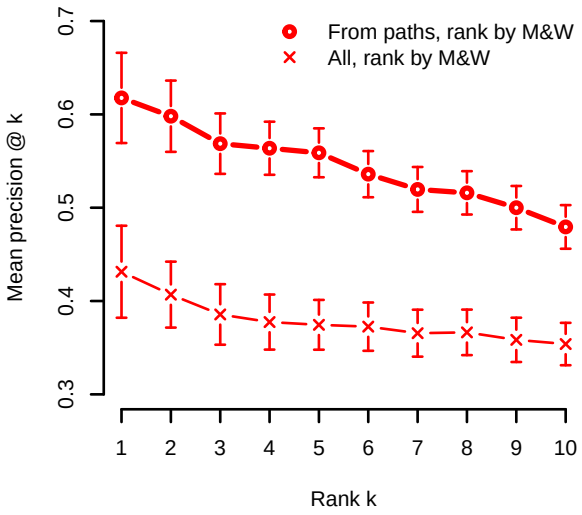

Figure 8: Precision@k with M&W ranking for human labeled ground truth.

We paid 5¢ per task, and each HIT was presented to 10 different workers. We consider a source as a positive label if over half of the 10 raters labeled it as such.

The task description given to the worker is reproduced *verbatim* in Appendix **??**.

As we observe from the results in Fig. 8, using path-based candidate selection followed by MW-ranking outperform MW-ranking on the set of all candidates by a large margin. Table 2, which summarizes the performance of both compared methods on the human-labeled data (again as the area under the precision@k curve) and compares it to the performance obtained on the automatically labeled data, shows that the area under the precision@k curves for path-based candidate selection increases by 12% points, compared to automatically obtained ground truth. On the other hand, precision when doing no candidate selection decreases by 1% point.

The automatically obtained ground truth uses only a historical notion of correctness and MW relatedness alone, without performing path-based candidate selection, might not capture the notion of human intuition-based similarity very well. Path-based selection, on the contrary, captures exactly that quantity by design, and it is not surprising that it prevails on a human-labeled ground truth by such a large margin.

Out of the two of the controls that represent randomly selected sources that already link to the target page (item 2 in the above list of source-candidate types presented to raters), only 9% are labeled as positive by at least 5 of the 10 raters. This tells us that the links we suggest are better than the average pre-existing link to the target.

On the other hand, out of the other two control links, which were essentially two completely random pages, only one pair (GEOGRAPHY

Candidate selection	Automatic ground truth	Human-labeled ground truth
None	39%	38%
Path-based	43%	**55**%

Table 2: Area under the precision@k curve for MW ranking.

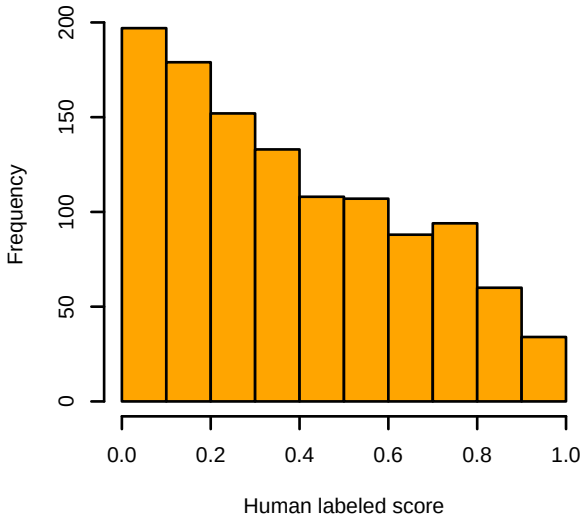

Figure 9: False negatives in the automatically obtained ground truth.

OF KOREA to SOUTH KOREA) was rated as a positive label by more than 5 of the 10 raters. This statistic testifies that human labeling was not random.

False negatives in the automatic ground truth. Now that we have human-labeled data, we can quantify the prevalence of false negatives in the automatically constructed ground truth. Consider Fig. 9, which shows a histogram of average human labels for the candidates that were labeled as negative according to the automatic ground truth. Here, 'average human label' refers to the average of the binary labels obtained from the 10 human raters for each candidate. We see that a large fraction of the examples labeled as negative according to the automatic ground are in fact positive examples, according to the more reliable human ground truth.

7. DISCUSSION

This paper introduces an effective method for the source prediction problem (Fig. 1(b)), in which a target page t is given, and the task is to find and rank sources s that should link to t. Prior work has primarily addressed the complementary target prediction problem (Fig. 1(a)), where s is given and t to be found. We consider source prediction more challenging than target prediction, since in the latter the set of link candidates is immediately given by the elements contained in the source page s (such as phrases or images), whereas in the former the set of source candidates must first be retrieved in a candidate selection step.

Computational feasibility. To illustrate this point, we briefly report on an experiment we had initially planned on doing. We intended to compare the performance of our method to the link predictions made by Milne & Witten's [9] machine-learned target prediction algorithm, but this was computationally infeasible: In order to use this target prediction method in a source prediction setting, we first had to find all articles s mentioning t (this required a full scan of a 44GB of Wikipedia dump). Next, we intended to annotate each source s with outgoing links and then rank s according to the score it gives to t. However, each annotation takes on the order of several seconds [**?**], but nearly every article mentions at least one of the targets we want to evaluate, so we would have had to annotate essentially all of Wikipedia, which would have taken a multiple of 3 million seconds, or several thousands of hours. One reason for the computational complexity of Milne & Witten's algorithm is that they (as well as other target prediction methods [6, 7, 18]) tend to spend significant effort on mention disambiguation.

On the contrary, in our approach, we neither have to scan Wikipedia for articles that mention t, nor do we need to do any sophisticated disambiguation or ranking. We simply use as source candidates all pages seen in our set of navigation traces, look for mentions only in this small subset of all Wikipedia pages, and rank according to a simple precomputed metric. This is possible because the brunt of the computational effort is done by humans: since they actively seek out pages that are likely to link to the target, these pages tend to be good source candidates, and issues such as disambiguation are much less critical.

Applications beyond Wikipedia. Now, we address the question if and how our technique could apply beyond the realm of Wikipedia. We envision two ways forward.

The first idea would be to gamify arbitrary websites. One could imagine a framework, e.g., written in JavaScript, that would wrap the website of interest, recruit players, and ask them to navigate to the targets we are interested in linking to. This would require adding at least some initial links to t manually, such that t is reachable by navigating. Furthermore, our method for finding valid anchors for the target, which is currently based on anchor-text/target-page pairs mined from Wikipedia Sec. 4, would need to be adapted to the new domain. Possibilities would include the use of prevalent phrases from the target's title and content as anchor texts, or, akin to our current method, the use of anchor texts that are already being used in other pages to refer to the target.

The second approach we envision is to use passively rather than actively collected log data for source-candidate selection and ranking. It might be possible to simply use the logs that are kept by webservers anyway. The added challenge here would be that we do not know what target (if any) a user tried to reach, whereas the target is always given explicitly to the user in the human-computation setup. However, we believe that reasoning along the following lines might be promising: If users that ended up in t often went through s, then the shortcut from s to t might be promising. An alternative heuristic might be to collect instances where a user navigates to s, issues a keyword query into the website's search box (if it exists), and clicks to t from the search-engine result page. We are currently working with the Wikimedia Foundation on a project that uses such passively collected log data for the purpose of missing-link mining.

Generalization to web graph. On Wikipedia, the linking guidelines are explicitly stated [16], so links are fairly consistent. Further, each page is typically about a single, well-defined topic. These are among the reasons why machine learning methods can infer powerful models for linking to Wikipedia articles. Websites other than Wikipedia are less likely to have the above properties, so it will be more difficult for statistical models to predict meaningful links. We expect methods for mining missing links directly from navigational traces to suffer less from this problem, since they do not take the detour through modeling the static structure of the link graph, but instead directly optimize navigability as the objective.

What we find especially promising in this light is a result from Fig. 6, namely that our method does not crucially rely on any measure of relatedness between pages: ranking our source candidates simply by the frequency with which they occurred in navigational traces for the given target (the yellow curve of Fig. 6) constitutes a competitive method. We believe that this makes our method a strong candidate for the source prediction task on websites other than Wikipedia, where a notion of relatedness between pages might be much harder to obtain.

APPENDIX

Here we give the description of the evaluation task given to human evaluators:

"Here's the deal! Our good friend Wikipedia is having self-doubts and wants you to help improve its links.

You are given a Wikipedia article (referred to as the target) and a list of other Wikipedia articles (referred to as source articles). You have to tell Wikipedia if the source article should contain a link to the target. And of course, if you are unsure of what the source or target article means, you can always click on the article name to open it in a new tab.

But remember that Wikipedia is a sensitive fellow and will be mad if you don't play by the rules: There should be a link from the source to the target if and only if

1. *the target article has some relevant information about the source article and could help readers understand the source article more fully*

2. *or the target article describes a proper name which is likely to be unfamiliar to readers."*

A. REFERENCES

[1] Wikipedia for schools. website, 2008.

[2] M. Ageev, Q. Guo, D. Lagun, and E. Agichtein. Find it if you can: A game for modeling different types of Web search success using interaction data. In *SIGIR*, 2011.

[3] A. Clemesha. The Wiki Game. Website, 2009. `http://www.thewikigame.com` (accessed Nov. 10, 2014).

[4] A. Clemesha. The wiki game, 2013.

[5] S. Fissaha Adafre and M. de Rijke. Discovering missing links in Wikipedia. In *LinkKDD*, 2005.

[6] E. Meij, W. Weerkamp, and M. de Rijke. Adding semantics to microblog posts. WSDM, 2012.

[7] R. Mihalcea and A. Csomai. Wikify! Linking documents to encyclopedic knowledge. In *CIKM*, 2007.

[8] D. Milne and I. H. Witten. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *WIKIAI*, 2008.

[9] D. Milne and I. H. Witten. Learning to link with Wikipedia. In *CIKM*, 2008.

[10] D. Milne and I. H. Witten. An open-source toolkit for mining wikipedia. *Artificial Intelligence*, 194:222–239, 2013.

[11] T. Noraset, C. Bhagavatula, and D. Downey. Adding high-precision links to wikipedia. 2014.

[12] R. West. Wikispeedia. website, 2009.

[13] R. West. Wikispeedia. Website, 2009. `http://www.wikispeedia.net` (accessed Nov. 10, 2014).

[14] R. West and J. Leskovec. Human wayfinding in information networks. In *WWW*, 2012.

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

[16] R. West, D. Precup, and J. Pineau. Completing Wikipedia's hyperlink structure through dimensionality reduction. In *CIKM*, 2009.

[17] Wikipedia. Wikipedia:manual of style/linking — wikipedia, the free encyclopedia. Website, 2014. `http://en.wikipedia.org/w/index.php?title=Wikipedia:Manual_of_Style/Linking&oldid=632571069` (accessed Nov. 10, 2014).

[18] Wikipedia. Wikiracing — wikipedia, the free encyclopedia. Website, 2014. `http://en.wikipedia.org/w/index.php?title=Wikiracing&oldid=630702489` (accessed Nov. 10, 2014).

[19] F. Wu and D. S. Weld. Autonomously semantifying Wikipedia. In *CIKM*, 2007.