# Discovering Missing Links in Wikipedia

Sisay Fissaha Adafre Maarten de Rijke
Informatics Institute, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
sfissaha,mdr@science.uva.nl

## ABSTRACT

In this paper we address the problem of discovering missing hypertext links in Wikipedia. The method we propose consists of two steps: first, we compute a cluster of highly similar pages around a given page, and then we identify candidate links from those similar pages that might be missing on the given page. The main innovation is in the algorithm that we use for identifying similar pages, LTRank, which ranks pages using co-citation and page title information. Both LTRank and the link discovery method are manually evaluated and show acceptable results, especially given the simplicity of the methods and conservativeness of the evaluation criteria.

## Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*clustering*; H.3.7 [**Information Storage and Retrieval**]: Digital Libraries—*systems issues*

## General Terms

Algorithms, Experimentation

## Keywords

Link analysis, Wikipedia, co-citation

## 1. INTRODUCTION

Wikipedia, the free on-line encyclopedia, is a hypertext document with a rich link structure [17]. Though its size is very small compared to the web, its link structure shares several properties with the web. For example, some Wikipedia pages, such as pages for countries like the USA or events like World War II, are cited more often than others resulting in a skewed distribution for the incoming and outgoing links, which is a typical characteristic of the web. On the web, the motivation for creating a hyperlink tends to vary, ranging from elaboration to referential to navigational to search engine optimization. Unlike the web, most hyperlinks in Wikipedia have a more consistent and semantically meaningful interpretation and purpose. For example, in Wikipedia hyperlinks to pages for the USA or World War II are created not because of their mere popularity but rather because of their close semantic relation with the page from which they link. While some of the links in Wikipedia are navigational, the majority is conceptual rather than navigational, often providing hierarchical information (showing parent-child relationships) or pointing to more detailed descriptions or definitions of the concept denoted by the anchor text, which form integral part of the content of the page. As a result, Wikipedia may be viewed as a semantic network-like structure, where the anchor texts denote the concepts and the links represent conceptual links.

Except for general formatting guidelines and a checklist [16], there are no strict rules for editing the content of Wikipedia. Authors can freely change the content of a Wikipedia page, and adding an outgoing link to a Wikipedia page is next to trivial: by including a term in double square brackets ([[`term`]]) one creates a link to a page with `term` as its ID or title. Group consensus is the main decision making process in determining the content and outgoing link structure of a Wikipedia page. While Wikipedia's lack of strict editorial guidelines leads to surprisingly few noisy links (as far as we are able to tell from anecdotal evidence), we do see evidence for another type of problem: *missing links*. By this we mean that valuable hypertext links are missing, or that hypertext links that are present do not have the best possible anchor text (and therefore target). E.g., one would expect that

- "Kenneth Arrow . . . is an [[American]] [[economist]] . . . ",
- "Lawrence Henry Summers . . . is an American [[economist]] . . . ",
- "Gary Stanley Becker . . . is an [[American]] economist . . . ",

which refer to three American economists would all have the same link structure as the first one, i.e., "X . . . is an [[American]] [[economist]] . . . ". Consistency at this level in the link structure is obviously important for readers, but it also matters for various Wikipedia-related lexical and information extraction tasks [1]. It may also improve results of other link-based analysis techniques.

How bad is the missing links problem in Wikipedia? Studies have shown that there are significant differences in the


*LinkKDD-2005,* August 21, 2005, Chicago, IL, USA.

links manually assigned by different people in hypertext documents [7]. We selected a sample of 44 professional tennis players' pages from Wikipedia and examined their link structures. Though the tennis players are expected to exhibit idiosyncratic link structure depending on their country of origin, achievements, etc., we expect that they all link to the concept "Tennis", given the fact that the word "tennis" is mentioned early in their respective Wikipedia pages, mostly in the first sentence. In Wikipedia, the first paragraph usually provides a very concise description of the entity, sometimes serving as definitions. However, only 32 out of the selected 44 are linked to the Wikipedia page for "Tennis". Similarly, out of 65 randomly selected singers, only 34 are linked to the concept "singer".

Our focus in this paper is on discovering such missing links. One key challenge is the semantic ambiguity of words or phrases, which implies that not all occurrences of a word or a phrase refer to the same concept defined in Wikipedia. Furthermore, not all links have equal importance or weight. Although a term appearing in a particular page refers to a concept defined in Wikipedia, it may have little or no conceptual relation with concepts being discussed in the page. The approach adopted in this paper makes use of a clustering technique to identify related entities and search for missing links in these pages. Specifically, to create clusters of similar pages for a given page $d$, we use a two-step ranking mechanism, LTRank, that exploits co-citation information as well as anchor text. One of the main requirements on LTRank is that it should find pages that are not just similar to the given page $d$, but that if $d$ is about a certain type of entity, then the similar pages should be about entities of the same type. E.g., if we attempt to discover links that might be missing from a page about a certain tennis player then LTRank should ideally only label pages of other tennis players as being similar.

Given a page $d$, and the pages most similar to $d$ according to LTRank, we extract suggestions for links missing from $d$ from the similar pages; the final step, then, is to identify links that are actually missing from $d$.

The remainder of the paper is organized as follows. In Section 2 we describe relevant features of our corpus, Wikipedia. Section 3 contains an overview of our approach to discovering missing links in Wikipedia, and in Sections 4 and 5 we detail the two key steps in our approach: given a Wikipedia page, find similar pages, and identify missing links. Section 6 contains a qualitative evaluation of the missing links we identify, while we discuss related work in Section 7 and conclude in Section 8.

# 2. ABOUT WIKIPEDIA

Wikipedia is a free online encyclopedia which is administered by the non-profit Wikimedia Foundation. The aim of the project to develop free encyclopedia for different languages. It is a collaborative effort of a community of volunteers, and its content can be edited by anyone. It is attracting increasing attention amongst web users and has joined the top 100 most popular sites.[1]

[1]See http://www.alexa.com/data/details/traffic_details?y=t&url=Wikipedia.org; site accessed on June 9, 2005.

## 2.1 The Wikipedia Corpus

As of May 16 2005, there are versions of Wikipedia in 200 languages. For the experiments in this paper, we used the English version of Wikipedia, which is the largest and contains more than 572k articles. We used the ascii text version of Wikipedia, which is available as database dump. Each entry of the encyclopedia (a page in the online version) corresponds to a single line in the text file. Each line consists of an ID (usually the name of the entity) followed by its description. The description part contains the body of the text that describes the entity. It contains a mixture of plain text and text with html tags. References to other Wikipedia pages in the text are marked using "[[" "]]" which corresponds to a hyperlink on the online version of Wikipedia. Most of the formatting information which is not relevant for the current task has been cleaned.

A Wikipedia page typically undergoes a number of revisions (on average 19.1 revisions per article for English) until a general consensus is reached among the authors regarding its content and outgoing links.

## 2.2 Link Structure

Wikipedia is a hypertext document with a rich link structure. As with typical web documents, Wikipedia pages differ in their content. The bulk of the Wikipedia pages provide a relatively complete description of an entity, they are *authorities* for their entities. Others act like focused *hubs*, providing a list of entities falling in a particular categories (such as list of male movie actors).

A description of an entity usually contains links to other pages within or outside of Wikipedia. The majority of these links correspond to entities, which are related to the entity being described, and have a separate entry in Wikipedia. As mentioned in Section 1, these links are used to guide the reader to a more detailed description of the concept denoted by the anchor text. This means that the links in Wikipedia typically indicate a topical association between the pages, or rather the entities described by the pages. E.g., in describing a particular person, reference will be made to such entities as country, organization and other important entities which are related to it and have entries in Wikipedia. In general, due to the peculiar characteristics of an encyclopedia corpus, the hyperlinks found in encyclopedia text are used to exemplify those instances of hyperlinks that exist among topically related entities [8, 13].

Each Wikipedia page is identified with a unique ID. These ids are formed by concatenating the words of the titles of the Wikipedia pages which are unique for each page, e.g., the page on Vincent van Gogh has "Vincent van Gogh" as its title and "Vincent_van_Gogh" as its ID. Each page may, however, be represented by different anchor texts in a hyperlink. The anchor texts may be simple morphological variants of the title such as plural form or may represent closely related semantic concept. For example, the anchor text "Dutch" points to the page for the Netherlands. In a sense, the IDs function as the canonical form for several related concepts. Although it may be difficult to say what the merits of the feature are for our task of discovering missing links, it definitely helps in minimizing the data sparseness problem, i.e., number of incoming links.

As regards the distribution of link counts, Wikipedia shows similar characteristics to the web. For example, the distribution of the incoming and outgoing links follow a power law [15].

## 3. OUR APPROACH

Our method for discovering links missing from a Wikipedia page consist of two main steps:

**Step 1** The first step concerns identification of topically related pages, i.e., clustering.

**Step 2** The second step involves identification of missing links. We identify candidate missing links and filter them through the anchor texts.

Before we jump into the details of the link discovery method, let us briefly consider a naive alternative method: simply identify and link those words or phrases in the Wikipedia pages that have an independent entry in Wikipedia. Though this method is easy to implement, there are a number of shortcomings associated with it. First, there are a number of common English words such as the article "A" and pronoun "It" that have separate entries. We do not want to create a link for every occurrence of these words. Second, most words or phrases have multiple meanings of which only a few correspond to the meaning of the Wikipedia entry. This in turn requires a separate disambiguation module, which will complicate the task.

Clustering alleviates most of the problems associated with the naive method. First, it enables us to restrict the application of the link discovery method to a particular cluster. Since links among pages entail relatedness, we prefer to establish the new links only among topically related pages. As we will see, clustering enables us to achieve this goal with a reasonable level of accuracy. Furthermore, in topically clustered Wikipedia pages, we expect that anchor texts tend to have the same meaning and refer consistently to the Wikipedia page. Moreover, in situations where an incorrect phrase is chosen, its application will be limited to that particular cluster hence avoiding global propagation of error.

Given that our overall aim is to identify links which are missing from a given page, it may not be necessary to enumerate all similar pages to a given page. Our working assumption is that a few closely related pages may suffice to identify the most important links. Furthermore, the method can be repeated multiple times to improve the link structure.

## 4. STEP 1: CLUSTERING PAGES

Finding the most closely related pages for a given Wikipedia page is the first step in the identification of missing links in Wikipedia hypertext corpus. One can try to find similar pages based on content, link structure or a combination of these [4, 10]. Here, we choose link structure since the main goal of the current study is recovering important missing links, therefore it is logical to look for Wikipedia pages that are similar based on their link structures. Therefore, we compute the similarity between Wikipedia pages based on the information we get from the link structure. More specifically, given a Wikipedia page, we abstract from its contents and represent it by the citations it receives, i.e., by its incoming links. We will use co-citations to identify similar pages: *two pages are similar if they are co-cited by a third*, and we assume that the co-citation counts correlate with the strength of similarity.

**Round 1**

1. Given a Wikipedia page $d$, collect all titles of the pages that link to $d$. Let $d_t$ be the resulting bag of terms; we call this the full title representation of $d$.
2. Use Lucene to index all full representations $d_t$ obtained in the first step.
3. For each Wikipedia page $d$, feed its full title representation $d_t$ as a query against the index built in the second step, producing a ranked list $L_d = d_{t_1}, \ldots, d_{t_n}, \ldots$, of pages represented as titles of the pages that link to them.

**Round 2**

4. Given a Wikipedia page $d$, consider the titles $t_1, \ldots, t_{10}$ of the 10 pages that form the top in the list $L_d$ generated in the third step. Represent $d$ as the bag of terms $d_s = \{t_1, \ldots, t_{10}\}$; we call this the short title representation of $d$.
5. Use Lucene to index all short title representations $d_s$ obtained in the fourth step.
6. For each Wikipedia page $d$, feed its short title representation $d_s$ as a query against the index built in step five, producing a ranked list $L'$. Collect the top $N$ pages in $L'$ whose similarity score is above some threshold $\alpha$. (For the experiments in this paper, we took $N = 100$, $\alpha = 0.3$.)

**Figure 1: Finding similar pages using LTRank.**

One naive method of identifying similar pages for a given page using co-citation information is to simply enumerate all Wikipedia pages that share a certain number of co-citations with the given page. However, pairwise comparison of all Wikipedia pages to get the most similar pages is impractical. Instead, we apply a widely used similarity measure from information retrieval. We used our own version of the Lucene search engine, an open-source, high-performance, full-text search engine written in Java [2]; it implements both vector-space and language models. We proceed in two stages: we start by representing each Wikipedia page by the IDs (or titles) of the citing pages; the pages are then indexed by treating all IDs as terms, and then every page (thus represented) is fed to Lucene as a query, producing a ranked list of pages for every page. This turns out to generate potentially off-topic lists; to overcome this issue, we need to restrict the terms used to represent pages.

Let us make things more precise now. The method that we propose is called LTRank ("Ranking based on Links and Titles"). The main steps are summarized in Figure 1. For every Wikipedia page $d$ we create its *full title representation* $d_s$ by collecting all titles of pages that cite $d$. Index all full title representations. Next, each page (represented by its citations) is submitted to the search engine as a query

in order to retrieve pages that are similar to it based on their citations. This completes Round 1 (in Figure 1), and gives us, for each Wikipedia page, a ranked list of similar pages. In most cases, however, the top ranking pages are topically closely related to the query page. In some cases, especially if the query page has many citations, the resulting list may contain a long list of unrelated pages. For example, the Wikipedia page for "Tea" has 247 incoming links. The ranked list of related pages for "Tea" that is generated using the 247 incoming links as a query contains some closely related pages, especially at the top, e.g., Coffee, Rice, Caffeine, China, Oolong, India, Xtea, Black tea. Further down the list, however, other pages which have little to do with the concept "Tea" show up, mainly concepts and people from cryptography e.g., RC6 (Block Cipher), David Wheeler (Computer Scientist).

This suggests that we should include a filtering mechanism, which is what we do in Round 2 (in Figure 1). To provide some intuitions, assume we have two Wikipedia pages together with their corresponding set of similar pages. If the two pages are similar, we expect there to be a high overlap between their corresponding sets of similar pages. Though the scenario is a bit different, a similar idea underlies the computation of SimRank [9]. In the case of SimRank, two documents are similar if the documents that cite them are similar providing a recursive mechanism for computing similarity measure. In the case of LTRank, similarity is computed using IR techniques. Therefore, in order to implement the above idea, we come up with a more lenient representation of Wikipedia pages: for each page $d$, we use the titles of the top 10 ranking similar pages from the list $L_d$ obtained in Round 1; these representations are called *short title representations*. The choice of 10, though arbitrary, is an attempt to keep the balance between a concise representation for a page on the one hand, and coverage on the other hand. The short title representations are then indexed using Lucene, after which each short title representation is submitted to the search engine. The output is a ranked list of Wikipedia pages. Finally, the pages whose retrieval status value is above a certain threshold (we use 0.3) are kept, and if the list is long, we only retain the top $N$ (we use $N = 100$).

Examination of the LTRank's output for a sample of Wikipedia pages shows that the set of similar pages identified by LTRank are (topically) closely related to the associated Wikipedia page. The list of similar pages may contain a set of homogenous items such as a list of "Tennis Players." In other cases, the list of similar pages may contain heterogeneous pages that fall under some broad concept, e.g., "Programmers," "Programming Concepts," and "Programming Languages" may fall under the broad concept "Programming," though they may look heterogeneous. For example, the list of similar pages for the Wikipedia page "Bertrand Meyer" (a programmer) consists mainly of programming concepts, rather than a set of programmers. This is sufficient for the current task which requires mainly of topical similarity: if a term is significant in one of the Wikipedia pages describing a programming concept, then we expect that the same term will be relevant if it occurs on the page for "Bertrand Meyer", a programmer.

To try and get a clearer idea of the qualitative performance of LTRank, a sample of the output for 20 pages has been selected and manually examined. We see two obvious ways to assess the output of LTRank: one is to take into account the natural semantic category to which (the topic of) a given page belongs and to demand that all similar pages found by LTRank belong to the category; the other is to simply demand that the pages returned by LTRank are relevant. Clearly, the former is more strict than the latter. For example, for "Andre Agassi" the obvious semantic category is *tennis player*, and according to the strict evaluation, similar pages should be pages of other tennis players, while the more lenient assessment criterion would also allow other types of entities, such as tennis championships that are held in different countries, such US Open, French Open etc., as long as they are relevant.

In Table 1, *Entity* refers to the Wikipedia page for which we are assessing the similar pages returned by our algorithm, *Category* refers to the category assigned to the page which is derived from the Wikipedia category information, *C-Similar* indicates the fraction of pages that fall under the category mentioned, and *Similar* is the fraction of pages that are found to be similar (without restriction to the category).

| Entity | Category | C-Similar | Similar |
|---|---|---|---|
| Andre Agassi | Tennis Player | 0.70 | 1.00 |
| Molar Gas Constant | Thermodynamics | 0.71 | 1.00 |
| Etienne Ys | Prime Minister | 0.00 | 0.86 |
| Nine Men's Morris | Game | 1.00 | - |
| Bertrand Meyer | Programmer | 0.25 | 1.00 |
| Power Kite | Kite | 0.75 | 0.75 |
| Shiloh | Biblical Places | 0.53 | 1.00 |
| Marilu Henner | Actors | 1.00 | - |
| City of Darebin | Australian City | 0.88 | 0.94 |
| Saa | Egyptian Mythology | 0.94 | 0.98 |
| Jacques Ruhlmann | Designers | 0.71 | 0.78 |
| Kusudama | Origami | 0.60 | 0.80 |
| Kirkwood, Missouri | town | 0.88 | 0.96 |
| West European Time | Time Zone | 0.64 | 0.64 |
| Drill n bass | Electronic Music | 0.68 | 0.95 |
| Manuel Jos de Arriaga | Politician | 0.59 | 1.00 |
| Third | Music | 0.97 | 1.00 |
| Legal Code | Ethics | 0.80 | 1.00 |
| Alexis Arguello | Boxer | 0.69 | 0.94 |
| Some Kind of Wonderful | Film | 0.92 | 1.00 |

**Table 1: Evaluation of 20 sample clusters identified by LTRank. The average cluster size was 28 documents, the minimum size 5 documents, and the maximum size 89 documents.**

For the first entity listed in Table 1, "Andre Agassi", 70% of the pages in the similar-page list are tennis players, and the rest are different tennis tournaments. In the case of "Etienne Ys", a prime minister of the Netherlands Antilles, the similar pages list contains no prime ministers but islands in the Caribbean Sea. The situation is similar for other instances shown in the table. For the current task, however, the pages on the similar-page lists are mostly relevant, as the numbers in the column labeled *Similar* indicate.

## 5. STEP 2: IDENTIFYING LINKS

Once we have identified the set of pages similar to a given page using LTRank, the next step is to search for important links missing from the given page. The search for missing

links is confined to this set of similar pages: our working hypothesis is that similar pages should have similar link structure.

We proceed as follows. We refer to the page that we analyze for missing links as the *main page* and to the pages identified as similar to the main page as *related pages*. Given a main page, we repeat the following steps. We take one of its related pages, and identify all outgoing links found in the related page that are absent from the main page; we also record the anchor texts for such links. The anchor texts are then searched in the main page. If an anchor is found then a link is added.

Although the anchor texts that we extract from the related pages may have different surface realizations on the main page (due to morphological variations, etc), the method works well for most instances since the set of similar pages usually contains multiple pages, which increases the likelihood of finding different surface realizations.

In general, the method identified missing links for 144,211 Wikipedia pages though not all are genuine missing links as is shown in the evaluation section. Table 2 provides some summary statistics regarding the output of the method.

| Proposed Missing Links | | |
|---|---|---|
| Min | Max | Average per Page |
| 1 | 132 | 4 |
| Outgoing links | | |
| Before | After | Overlap |
| 27 | 32 | 0.16 |

**Table 2: Summary statistics on outgoing links and identified missing links.**

In Table 2, the upper part shows the minimum, maximum and average number of missing links (per page for which links were proposed) suggested by the system. The maximum number of missing links, i.e., 132, was proposed for the page *fascism*, which contains multiple terms in the area of politics that have entries in Wikipedia. The lower part of Table 2 shows the average number of outgoing links on these pages. *Before* refers to the average number of outgoing links without the identified missing links, and *After* refers to the average with the identified missing links. For certain pages, an existing link is identified as a missing link. *Overlap* refers to the average number of this kind of links per page. These errors are mainly due to the "Redirect" facility of Wikipedia that allows two anchor texts with different IDs to point to the same Wikipedia page as will be explained in the evaluation section.

## 6. QUALITATIVE EVALUATION

For the task we are addressing in this paper (discovering missing links), it is difficult to compute the typical evaluation measures, i.e., precision and recall, since we do not have an exhaustive list of all the missing links. Instead, we turn to sampling-based evaluation. We take a random sample of 100 links that are identified by the method and manually examined the links. The output of the method described in Section 5 is a list of proposed links. Each entry in the list consists of an anchor text of the new link, the Wikipedia page containing the anchor text, and a similar page which has the link with this anchor text.

DIRECT3D is part of [[Microsoft]]'s [[DirectX]] [[API]]. Direct3D is only for use in Microsoft's various [[Microsoft Windows—Windows]] [[operating systems]] ([[Windows 95]] and above) and, although in a quite different version, in the [[Xbox]]. Direct3D is used to render [[3D computer graphics—three dimensional graphics]] in applications where top performance is important, such as games. Direct3D also allows applications to run fullscreen instead of embedded in a window, though they can still run in a window if programmed for that feature. Direct3D uses [[hardware acceleration]] if it is available on the graphic board.

Direct3D is a [[3D computer graphics—3D]] API. That is, it contains many commands for 3D **rendering**, but contains few commands for rendering [[2D computer graphics—2D]] graphics. Microsoft continually updates Direct3D with the latest technology available on 3D graphics cards. Direct3D offers full **vertex** software emulation but no **pixel** software emulation for features not available in hardware.
...

**Figure 2: Part of a Wikipedia page for Direct3D. Missing links suggested by our method are indicated in boldface.**

In order to have a consistent evaluation of the results, we used the following criteria. The first criterion checks if the anchor text has the same meaning as, or represents well, the Wikipedia page that it points to. Once the proposed link passes this test, we check if it is actually relevant; by necessity, this is a subjective decision. We examine various aspects of the proposed link in order to reduce the amount of subjectivity in determining the relevancy of a link. As described in Section 5, the link identification process involves two Wikipedia pages, i.e. a *main page* and a *related page*, which are identified as being topically related pages. Therefore, one criterion, which is less subjective, is to check if the two pages are, indeed, closely related. Otherwise, the proposed link will be based on poor evidence, and hence will be treated as noise. For example, if the pages are about presidents of two countries and the proposed link points at the page for "President", it is highly likely that the new link is relevant.

Using the above criteria, out of the sample of 100 proposed links 68 are found to be relevant. One piece of evidence that suggests that our link discovery method has to a large extent achieved its goal, is that among the links which are labelled as noisy none has been found to be caused by ambiguous anchor texts. Figure 2, which shows a portion of the Wikipedia page for "Direct3D", exemplifies the sort of links that have been identified by the method. For this particular example, the method proposed three additional links on the basis of the evidence obtained from the following three related pages, i.e., "Graphics Card", "Scene Graph", "Vertex and Pixel Shader." These are links to "Pixel," "Vertex,"

RICHARD BURNS is a [[rally]] driver from [[England]]. He was born [[January 17]] [[1971]] at the [[Royal Berkshire Hospital]],[[Reading, England]].

He started driving in field near his house at the tender age of 8 in his fathers old [[Triumph Motor Company—Triumph]] 2000. At 11 Richard joined an under 17's car club, where he became driver of the year in 1984.

Just two years later his father arranged a trip to Jan Churchill's [[Wales—Welsh]] Forest Rally School near [[Newtown]] where Richard drove a [[Ford Escort]] for the day, from that moment on Richard knew what he wanted to do.

Richard badgered his father into letting him join the [[Craven Motor Club]] in his home **town** Reading where his talent was spotted by rally raconteur and enthusiast [[David Williams]] and where he rallied the stages of Panaround, Bagshot, Mid-Wales, Millbrook, Severn Valley, Kayel Graphics and Cambrian Rally.
...

**Figure 3: Part of the Wikipedia page for Richard Burns. A missing link suggested by our method is indicated in boldface.**

and "Rendering" pages. All three concepts are closely related to the theme of "Direct3D." Though it might be said that the first two may be subsumed by other links in the page such as "Pixel Shaders" and "Vertex Shaders", "Rendering" seems to require a link as it is often mentioned in the page signifying its importance.

Some of the proposed links are labelled noisy due to a lack of sufficient evidence for their being relevant in the specified context. For example, a link (with anchor text "town") from the "Richard Burns" page to "town" was proposed based on the result of comparing the link structure of "Richard Burns" with its similar page "Earley" (town). Both happen to be from the same county and share some links, which explains the similarity. However, their similarity is not sufficient enough to warrant the sort of inference we are trying to make. In case of "Earley", the link "town" is mentioned as part of its definition whereas in the case of "Richard Burns" the choice of the word "town" is rather random and hence less prominent, i.e., the same information could have been expressed using other lexical items as is shown in Figure 3. Another kind of error stems from Wikipedia's redirection facility. This facility enables one to link two anchor texts with different Wikipedia IDs to the same Wikipedia pages. This feature is typically used to redirect abbreviations to the pages of their extended versions. However, it becomes a problem when the same anchor text gets two distinct IDs. Since we are using the IDs, the occurrences will be treated as different. As a result, a new link has been incorrectly suggested although the link already exists. An example of this error is "Legislative" which is represented by two IDs, "Legislative" and "Legislatur".

# 7. RELATED WORK

We briefly discuss two types of related work: one concerns Wikipedia, the other concerns link analysis.

## 7.1 Wikipedia

Though Wikipedia is very young, it has grown to be the largest free online encyclopedia in a short-period of time. Currently it has reached a level where it can support different types of research, concerning multi-authored content creation, collaborative learning, and link structure. Ciffolilli [5] describes the type of Wikipedia's community, processes of reputation and reasons for its success. Viégas et al. [14] introduce a method for visualizing edit histories of Wikipedia pages and found some collaboration patterns. Lih [11] analyzes citations of Wikipedia articles in the press and the ratio between number of edits and unique editors. Miller [12] deals with the blurring distinction between reader and author. One example of Wikipedia related research that is directly relevant to the present paper is due to Bellomi and Bonato [3], who applied link-based analysis to compute lexical authorities. Finally, Voss [15] provides a general description of Wikipedia and a broad review of Wikipedia related research.

## 7.2 Link analysis

In contrast to research on Wikipedia, link analysis is a relatively mature discipline with an extensive literature on the topic. One of the applications of link-based analysis techniques is identification of topically related items. Document clustering typically makes use of similarity measures that are based on terms derived from the content of the document. With the coming of the web and hypertext documents, link based techniques are gaining popularity. One prevailing assumption is that the link between entities implies certain degree of relatedness, and the link density correlates with the degree of relatedness. On the basis of these assumptions, a number of link-based techniques have been developed and applied to solve diverse problems. One of the most widely cited applications of link-based analysis techniques is to improve search engine results [4]. Traditional IR systems rely on content-based analysis techniques in order to retrieve and rank documents. With the emergence of the Web and hypertext documents, however, the relevance of a document is not solely measured by its content only but also based on the structural context which it finds itself in.

Typically, identification of a cluster involves searching for certain graph structures. Co-citation and bibliographic coupling are the two link based similarity measures. In the case of co-citation, two pages are related if they are co-cited by a third document. The strength of the relation is usually measured by the number of co-citation counts. In the case of bibliographic coupling, on the other hand, two pages are related if both cite the same document. In this paper we used co-citation ideas in developing LTRank.

Kumar et al. [10] used the concept of co-citation and graph-based techniques to identify emerging Web communities with members that are topically related. The basic idea is to identify dense bipartite graphs of small sizes and use them to identify other members of the community. Similarity is based on simple co-citation counts.

SimRank is another structural similarity measure which is based on the 'random surfer' model [9]. The intuition un-

derlying SimRank is that two objects are similar if they appear in a similar structural context. In case of a web graph, two pages are similar if they are linked to by similar pages. This formulation entails recursive computation of similarity scores. As it is impossible to compute similarity measures for all possible pairs (large $n$), only node pairs within certain radius apart are considered (2, 3). As pointed out in Section 4, LTRank shares a number of intuitions with SimRank. Finally, Companion is another neighborhood-based algorithm for finding related pages. For a given page, it identifies a set of closely located neighborhoods and computes authority scores for these pages and returns the highest ranking page as the most closely related page [6].

## 8. CONCLUSION

In this paper we addressed the problem of discovering missing links in Wikipedia. The method we proposed consists of two steps: first identify a cluster of highly similar pages around a given page, and then identify candidate links from the similar pages that might be missing on the given page. The main innovation is in the algorithm that we used for identifying similar pages, LTRank, which ranks pages using co-citation and page title information. Both LTRank and the discovery method were evaluated and showed acceptable results, especially given the simplicity of the methods and conservativeness of the evaluation criteria. Though the methods are not perfect, they could be used as an online authoring aid by revealing a ranked list of important candidate links, and the associated Wikipedia links. To some extent, this would provide a page's author with a global view of the structure of Wikipedia while locally updating or editing a page.

As to future work, we believe that there is a particularly appealing property that we get from the use of a search engine in LTRank, and that we want to explore further, i.e., term weighting. The terms, in our case, are IDs of incoming links. The search engine we use in LTRank, Lucene, uses tf.idf term weighting. The advantage of weighted incoming links becomes clear when we look at the correlation among the number of links in a page, the size of a page, and also diversity of topics. Wikipedia pages that have several outgoing links tend to be longer (on average there is about 1 hyperlink for every 17 words), and usually deal with diverse topics. In contrast, Wikipedia pages with few outgoing links tend to be shorter and homogeneous in terms of the topics dealt with. What this suggests is that there is a higher likelihood that pages that are co-cited by a page with several outgoing links are less similar than those co-cited by the short pages or pages with few outgoing links. Lucene enables us to capture these intuitions by assigning less weight to citations coming from longer pages. A brief examination of the output of LTRank supports this hypothesis, although it needs to be tested more thoroughly.

Furthermore, though we were not able to carry out a detailed theoretical comparison of LTRank method with other related techniques due to time constraints, the use of a well established information retieval technique in finding similar pages seems to add to the effiency of computation of similarity scores since it avoids pairwise comparison of Wikipedia pages (as SimRank would require). However, this should be empirically tested.

So far, we only considered co-citation as a basis for our similarity measure. However, it may be useful to take a broader view of the neighborhoods of a page, which also includes the pages cited, and search for related pages based on the extended set. This in turn would allow one to take into account the properties of another important similarity measure, bibliographic coupling, i.e., two documents are similar if they cite the same document. As mentioned previously, there is a high degree of overlap between the content and outgoing links in the Wikipedia corpus. Therefore, searching for similar pages by making use of outgoing links also partially overlaps with searching using the content of the Wikipedia pages. In addition, since we are looking for *missing* outgoing links, it may also be natural to search for pages that are similar in terms of their outgoing link pattern.

Finally, it is natural to ask whether the techniques developed in this paper can be applied in other contexts, particularly on a corpus obtained from the web. In order to shed some light into this issue, it may be worthwhile to recall those properties of Wikipedia which may have contributed (a lot) towards the results we achieved. As mentioned in previous sections, the semantic-network like nature of the Wikipedia hyperlink structure seems to be its characteristic feature which somehow distinguishes it form the web. This property in turn results in a relatively dense network structure in which the edges indicate important semantic relationships. The fact that the network is dense to some extent entails sufficiently large sets of recurrent patterns, which is important for a method based on co-citation counts. Furthermore, as in any other encyclopedia, there seems to be some kind of redundancy in Wikipedia, as certain classes of entities tend to be covered extensively. Though the content of Wikipedia can be edited by anyone, its quality is controlled indirectly through group consensus. This in turn ensures good quality content, anchor texts, and network structure. In contrast, the web tends to be more noisy without any control on its content and structure. Though some of these problems may be compensated through the mere size of the web, it may not be sufficient to match the advantages one gets from a corpus developed in a controlled environment.

## 9. ACKNOWLEDGMENTS

Sisay Fissaha Adafre was supported by the Netherlands Organization for Scientific Research (NWO) under project number 220-80-001. Maarten de Rijke was supported by grants from NWO, under project numbers 365-20-005, 612.069.006, 220-80-001, 612.000.106, 612.000.207, 612.066.302, 264-70-050, and 017.001.190.

## 10. REFERENCES

[1] D. Ahn, V. Jijkoun, G. Mishne, K. Müller, M. de Rijke, and S. Schlobach. Using Wikipedia at the TREC QA Track. In *Proceedings TREC 2004*, 2005.

[2] Apache Lucene. A high-performance, full-featured text search engine library. URL: http://lucene.apache.org, 2005.

[3] F. Bellomi and R. Bonato. Lexical authorities in an encyclopedic corpus: a case study with wikipedia. URL: http://www.fran.it/blog/2005/01/lexical-authorities-in-encyclopedic.html, 2005. Site accessed on June 9, 2005.

[4] S. Chakrabarti. *Mining the Web*. Morgan Kaufmann, 2002.

[5] A. Ciffolilli. Phantom authority, selfselective recruitment and retention of members in virtual communities: The case of Wikipedia. *First Monday*, 8(12), 2003.

[6] J. Dean and M. R. Henzinger. Finding related pages in the world wide web. *Computer Networks (Amsterdam, Netherlands: 1999)*, 31(11–16):1467–1479, 1999.

[7] D. Ellis, J. Furner-Hines, and P. Willett. On the measurement of inter-linker consistency and retrieval effectiveness in hypertext databases. In *SIGIR 1994: Proceedings of the 17th Annual International Conference on Research and Development in Information Retrieval*, pages 51–60, 1994.

[8] R. Ghani, S. Slattery, and Y. Yang. Hypertext categorization using hyperlink patterns and meta data. In C. Brodley and A. Danyluk, editors, *Proceedings of ICML-01, 18th International Conference on Machine Learning*, pages 178–185, 2001.

[9] G. Jeh and J. Widom. SimRank: a measure of structural-context similarity. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 538–543, 2002.

[10] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the Web for emerging cybercommunities. *Computer Networks*, 31(11–16):1481–1493, 1999.

[11] A. Lih. Wikipedia as participatory journalism: Reliable sources? Metrics for evaluating collaborative media as a news resource. In *Proceedings of the 5th International Symposium on Online Journalism*, 2004.

[12] N. Miller. Wikipedia and the disappearing "Author". *ETC: A Review of General Semantics*, 62(1):37–40, 2005.

[13] U. Rao and M. Turoff. Hypertext functionality: A theoretical framework. *International Journal of Human-Computer Interaction*, 1990.

[14] F. Viégas, M. Wattenberg, and D. Kushal. Studying cooperation and conflict between authors with history flow visualization. In *Proceedings of the 2004 conference on Human factors in computing systems*, 2004.

[15] J. Voss. Measuring Wikipedia. In *Proceedings 10th International Conference of the International Society for Scientometrics and Informetrics*, 2005.

[16] Wikipedia. Manual of style. URL: http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style_%28links%29, 2005.

[17] Wikipedia. The Free Encyclopedia, 2005. URL: http://www.wikipedia.org.