# Topical and Structural Linkage in Wikipedia

Kelly Y. Itakura[1,2], Charles L. A. Clarke[1], Shlomo Geva[2], Andrew Trotman[3], and Wei Chi Huang[2]

[1] School of Computer Science, University of Waterloo, Canada
[2] School of Information Technology, Queensland University of Technology, Australia
[3] Department of Computer Science, University of Otago, New Zealand

**Abstract.** We explore statistical properties of links within Wikipedia. We demonstrate that a simple algorithm can predict many of the links that would normally be added to a new article, without considering the topic of the article itself. We then explore a variant of topic-oriented PageRank, which can effectively identify topical links within existing articles, when compared with manual judgments of their topical relevance. Based on these results, we suggest that linkages within Wikipedia arise from a combination of structural requirements and topical relationships.

## 1 Introduction

The internal link structure of Wikipedia differs substantially from the structure of the general Web. Understanding this structure may afford insights to the editors who develop its content, and may also assist the growing league of researchers employing Wikipedia as a convenient and general source of machine-usable knowledge regarding human language, society, history, science, and other subjects. In addition, we study links in Wikipedia with the aim of automatically suggesting out-going links from new articles, and new links for existing articles. We derive our inspiration from the INEX Link-the-Wiki track [4, 5], which includes a task to restore links to a set of Wikipedia articles stripped of these links.

At INEX 2007, task participants returned ranked lists of suggested links for each stripped article. By treating the original links appearing in the articles as ground truth, precision and recall measures were applied to evaluate the results. For the INEX 2007 task, we implemented a simple statistical approach that substantially outperformed other approaches [6]. We call this approach the *structural threshold algorithm*, or just the *ST algorithm*. We were surprised by the effectiveness of the ST algorithm because it does not consider the topic of the article forming the source of the link, but only the anchor phrase and the target of the link. At INEX 2008, the ST algorithm was independently implemented by multiple participants, successfully demonstrating that the algorithm effectively recovers many of the links in the original articles.

Along with an evaluation against Wikipedia ground truth, the INEX 2008 evaluation process included separate manual judgments, in which assessors identified links that they believed were topically relevant to the articles. When runs

were measured against these manual judgments, it transpired that both the submitted runs and the original Wikipedia ground truth differed substantially from these manual judgments. Based on this experience, we hypothesize that many links in Wikipedia are primarily *structural* — that many anchor phrases are frequently linked to associated articles regardless of the topic of the linking article. On the other hand, some links are clearly *topical* in nature, created to express a specific relationship between a pair of articles.

## 2 Related work

A small body of prior work has addressed the problem of link discovery in Wikipedia. This work often employs statistical similarity measures [8, 9] or cocitations [1] to compute the probability that a phrase constitutes an anchor. Similar to the work of Mihalcea and Csomai [8] and the work of Milne and Witten [9], our ST algorithm depends on simple statistics. Like the work of Mihalcea and Csomai, the ST algorithm ranks anchor phrases by link probability, but unlike that work, the ST algorithm considers more than just article titles as potential anchor phrases. On the other hand, Milne and Witten, build on top of an ST-like algorithm by using the textual context of an article to disambiguate anchor phrases and destination pages. Gardner and Xiong [3] employ a sequence labeling approach, in which a label indicates the start and part of the links. They use a conditional random field as a classifier, training on both local and global data. Their results are comparable to those of Milne and Witten.

Mihalcea and Csomai, Milne and Witten, and Gardner and Xiong all treat existing links in Wikipedia as manually-crafted *ground truth*. The ST algorithm's performance against this ground truth is better than that of Mihalcea and Csomai and somewhat worse than the more complicated approach of Milne and Witten. However, by applying a variant of PageRank, we demonstrate that existing Wikipedia links should not necessarily be regarded as a gold standard. Instead, topical and structural linkages should be considered separately; otherwise, structural linkages tend to dominate the evaluation.

## 3 Structural Threshold Algorithm

To suggest links for a source article, the structural threshold (ST) algorithm first computes link probability estimates for potential anchor phrases and then applies a cutoff based on anchor density. In applying the ST algorithm, we imagine that this source article was newly added to Wikipedia. In reality, for the purposes of our experiments, the source article was removed from our Wikipedia corpus and stripped of links. For all of our experiments, we work with the Wikipedia corpus used at INEX 2008.

### 3.1 Link Probability Estimate

The ST algorithm first creates a list of all potential anchor phrases by considering all phrases appearing the source article up to some fixed length, after whitespace

normalization. For each potential anchor phrase $a$, we assign the destinations $d$ in order of their frequency in Wikipedia. For each most frequent destination $d$, we compute the probability estimate $\gamma$ that a phrase will be linked as an anchor as follows:

$$\gamma = \frac{\sharp \text{ of pages containing a link from anchor } a \text{ to a destination page } d}{\sharp \text{ of pages in which } a \text{ appears at least once}} \ .$$

Links are suggested by the ST algorithm in order of decreasing $\gamma$ values. The ST algorithm essentially creates a ranked list of possible links from an article. Since Wikipedia articles do not link to all possible destinations, the actual list of links to add to an article might be determined by setting a threshold for $\gamma$ based on article length, as we discuss next.

### 3.2 Anchor Density

To select a threshold for adding links to an page we consider the overall density of anchors in an average article. We define an *anchor density* $\delta$ of a page $p$ as follows.

$$\delta_p = \frac{\#\text{ of article linked from page } p}{\text{size of page } p \text{ without tags in KB}} \ .$$

The average anchor density of a corpus with $N$ pages is, therefore,

$$\delta = \sum_p \frac{\delta_p}{N} \ .$$

We estimated overall anchor density for the corpus by allocating articles into 5KB bins, according to their size, and then computing the average for each of the bins. The density is roughly linear, with approximately 7.09 anchors for every KB [6]. Similar observations have been made by Zhang and Kamps [11]. Thus, instead of selecting links according to a $\gamma$ threshold, we may add links to an article in $\gamma$ order, until an anchor density of $\delta = 7.09/KB$ is reached.

### 3.3 Performance

INEX 2007 participants generated ranked lists of possible links for each article, which were evaluated against Wikipedia ground truth using standard recall and precision measures. Table 1 compares the performance of the ST algorithm against the best performance achieved by other approaches. As a result of this success at INEX 2007, several other INEX 2008 participants incorporated the ST algorithm into their efforts.

To provide another performance comparison, we incorporated anchor density into an INEX 2008 run that used the ST algorithm, cutting the list of anchor phrases ranked by $\gamma$ for each topic by $\delta$ times the article's file size. We evaluated the effect of $\delta$ by computing a set precision and recall against the Wikipedia ground truth. We obtained the precision of 62.88% and recall of 65.21%. Mihalcea and Cosmai report results of 53.37% precision and 55.90% recall on an

| | MAP | rPrec | P@5 | P@10 | P@20 |
|---|---|---|---|---|---|
| ST algorithm | 0.61 | 0.63 | 0.85 | 0.82 | 0.75 |
| Second-best run | 0.32 | 0.42 | 0.77 | 0.68 | 0.58 |

**Table 1.** Performance of the structural threshold algorithm at INEX 2007

equivalent task [8]. Milne and Witten, with a more complex anchor disambiguation technique, achieve precision of 74.4% and recall of 73.8% [9]. The simpler ST approach provide reasonable effectiveness despite its lack of any topical considerations.

## 4 Link Analysis

The ST algorithm identifies all anchor phrases with high $\gamma$ regardless of the topic of an article. Thus, the anchor phrase "Peninsular War", with $\gamma = 0.898$, would be suggested as an anchor for nearly any page in which it occurs, from "Napoleon" to "Otago" to "wine", regardless of its topical relationship to those pages. According to Wikipedia ground truth, "Peninsular War" should be indeed linked into the article on "Otago", a city in New Zealand. However, our manual assessors deemed the anchor phrase to be topically non-relevant; from an assessor's point of view, this European war is not related to New Zealand. On the other hand, the page "Otakou", from where "Otago" derived its name, appears to be topically relevant, and the manual assessors agree. However, it has a relatively low $\gamma$ value of 0.63 and the Wikipedia ground truth considers it non-relevant.

### 4.1 KPR Algorithm

We enlist a variant of PageRank to estimate the topicality of anchor phrases. Instead of measuring the popularity of anchor phrases by the $\gamma$ value, this variant balances both topicality and popularity by computing the contribution to KL-divergence of scores between a standard PageRank (PR) and a topic-oriented PageRank (TPR). Details of this algorithm, which we call the *KPR Algorithm*, are given as an example in Büttcher et al. [2, pages 526–528].

For the work reported in this paper, we use a process that computes KPR with all links present — no articles or links are removed. Our goal is to show that even though there are both topical and structural links in Wikipedia, the topical links indicated by KPR provide a better match to the manual judgments. For INEX 2009 (in work not reported here) we applied KPR to solve the link discovery problem [7]. In those experiments, we used a stripped Wikipedia corpus, employed the ST algorithm to provide initial linkages for the source article, and then applied KPR to compute the final linkages.

Table 2 lists the top-10 results ordered by KPR values for the source article "Otago". PR, TPR, and $\gamma$ values are included for comparison. Generally, the results appear to be closely related to the topic of "Otago".

| Article | Relevance | KPR | TPR | PR | $\gamma$ |
|---|---|---|---|---|---|
| Dunedin | 1 | 0.0171 | 4302 | 1.28 | 0.61 |
| Central Otago Gold Rush | 1 | 0.0166 | 3278 | 1.49 | 0.83 |
| South Otago | 1 | 0.0165 | 2962 | 0.62 | 0.44 |
| New Zealand | 1 | 0.0156 | 7524 | 321.33 | 0.68 |
| Otakou | 1 | 0.0151 | 2688 | 0.52 | 0.63 |
| Balclutha, New Zealand | 1 | 0.0151 | 279 | 0.77 | 0.55 |
| Gabriel Read | 1 | 0.0149 | 2643 | 0.52 | 0.71 |
| Invercargill | 0 | 0.0147 | 3299 | 3.84 | 0.80 |
| South Island | 1 | 0.0146 | 4259 | 23.37 | 0.79 |
| Queenstown, New Zealand | 1 | 0.0144 | 3007 | 2.12 | 0.78 |

**Table 2.** Top 10 results for "Otago" ranked by KPR compared with manual relevance values

### 4.2 Performance

The aim of the KPR algorithm is to identify articles that are topically related to a given source article. Since the assessors for INEX 2008 judged an article relevant only when it was topically related to a source article, we would expect that ranking links by the KPR algorithm would produce results closer to the manual judgments than ranking links by the ST algorithm. To test this hypothesis, we ranked INEX 2008 Wikipedia ground truth using both the KPR algorithm and the ST algorithm.

For each article in the INEX 2008 test set, we extracted its original links and applied both algorithms to these links. We computed P@5, P@10, and P@20 for each ranking against the manual assessments. The results are shown in Table 3. Even though these articles contain both topical links with high KPR values and structural links with high $\gamma$ values, the manual assessors prefer those with high KPR values.

If we view the manual assessments as the gold standard, our INEX experience indicates that comparisons against Wikipedia ground truth is not an ideal method for assessing the performance of link suggestion algorithms. Moreover, manual assessments are not easily scalable to experiments with thousands of topics. However, our experience with the KPR algorithm indicates that KPR is a reasonable indicator of topical relevance.

These observations lead to the idea of applying the KPR algorithm to automatically construct a topically oriented assessment set. To create this set, we first applied the KPR algorithm to articles from the INEX 2008 task. For each article, we took the top-10 links by KPR values and labeled them as topically relevant, as if they were judged that way by a human assessor. All other links were considered to be not topically relevant.

We use Kendall's $\tau$ to compare the rankings under the two assessment sets. While the resulting value of $\tau = 0.7354$ falls short of the level that might allow us to replace the manual assessments with automatic assessments [10], it suggests a close relationship between the two sets. Given the simplicity of this experiment,

| | P@5 | P@10 | P@20 |
|---|---|---|---|
| ST algorithm ($\gamma$) | 54.80 | 56.20 | 51.03 |
| KPR algorithm | 66.40 | 63.40 | 59.93 |
| $\Delta$ | 11.60 | 7.20 | 8.90 |
| $p$-value | 0.0007 | 0.0337 | 0.0003 |

**Table 3.** Wikipedia ground truth ranked by ST values ($\gamma$) and KPR values as measured against manual assessments

where the top-10 links are simply assumed to be relevant, additional development of this approach may produce a stronger correspondence.

## 5 Concluding Discussion

Links form the glue that binds Wikipedia together. As we demonstrate in this paper, these links appear to arise from a combination of structural requirements and topical relationships. Even in the absence of topical information, we can make reasonable predictions about the links appearing in an article. However, in order to predict manual assessments of topical relevance, the interconnections between articles must be considered.

## References


1. Adafre, S.F., de Rijke, M.: Discovering missing links in Wikipedia. In: 3rd International Workshop on Link Discovery. pp. 90–97. Chicago (2005)
2. Büttcher, S., Clarke, C.L.A., Cormack, G.V.: Information Retrieval: Implementing and Evaluating Search Engines. MIT Press, Cambridge, Massachusetts (2010)
3. Gardner, J.J., Xiong, L.: Automatic link detection: A sequence labeling approach. In: 18th CIKM. pp. 1701–1704. Hong Kong (2009)
4. Huang, D.W.C., Xu, Y., Trotman, A., Geva, S.: Overview of INEX 2007 Link the Wiki track. In: INEX 2007 (Springer LNCS 4862). pp. 373–387 (2007)
5. Huang, W.C., Geva, S., Trotman, A.: Overview of the INEX 2008 Link the Wiki track. In: INEX 2008 (Spring LNCS 5631). pp. 314–325 (2008)
6. Itakura, K.Y., Clarke, C.L.A.: University of Waterloo at INEX2007. In: INEX 2007 (Springer LNCS 4862). pp. 417–425 (2007)
7. Itakura, K.Y., Clarke, C.L.A.: University of Waterloo at INEX 2009: Ad Hoc, Book, Entity Ranking, and Link-the-Wiki tracks. In: INEX 2009 (Springer LNCS 6203). vol. 6203, pp. 331–341 (2010)
8. Mihalcea, R., Csomai, A.: Wikify!: Linking documents to encyclopedic knowledge. In: 16th CIKM. pp. 233–242. Lisbon (2007)
9. Milne, D., Witten, I.H.: Learning to link with Wikipedia. In: 17th CIKM. pp. 509–518. Napa Valley, California (2008)
10. Voorhees, E.M.: Variations in relevance judgments and the measurement of retrieval effectiveness. In: 21st SIGIR. pp. 315–323 (1998)
11. Zhang, J., Kamps, J.: Link detection in XML documents: What about repeated links. In: SIGIR 2008 Workshop on Focused Retrieval. pp. 59–66 (2008)