Link Discovery in the Wikipedia

Shlomo Geva [1], Andrew Trotman [2], Ling-Xiang Tang [1],

[1] Faculty of Science and Technology,
Queensland University of Technology,
Brisbane, Australia
{s.geva, l4.tang}@qut.edu.au
[2] Department of Computer Science,
University of Otago,
Dunedin, New Zealand
andrew@cs.otago.ac.nz

Abstract. In this paper we describe our approaches taken in the Link-the-Wiki track. We submitted runs for all three Link-the-Wiki tasks: Link-the-Wiki, Link-Te-Ara, and Link-Te-Ara-to-the-Wiki. To generate outgoing links for each task, our link discovery system employs the top ranking algorithms from previous LTW tracks and a hybrid method derived from them. For incoming links, we used traditional information retrieval strategy on the Wikipedia XML collection. The official results for the INEX 2009 Link-the-Wiki track show encouraging performance of our system.

Keywords: Wikipedia, Link Discovery, Best Entry Point.

1 Introduction

We submitted runs for all Link-The-Wiki tasks: Link-the-Wiki, Link-Te-Ara, and Link-Te-Ara-to-the-Wiki. The Link-the-Wiki task requires the identification of incoming links and outgoing links at both file-to-file and anchor-to-bep levels; the Link-Te-Ara requires the identification of anchor-to-bep outgoing links for all documents in the Te Ara Encyclopedia; and the Link-Te-Ara-to-the-Wiki requires the identification of anchors in Te Ara pages and their corresponding BEPs in the Wikipedia corpus.

The algorithms we used to generate the outgoing links for each task are based on the top-ranking link generation algorithms from previous years: Itakura's algorithm; and Geva's algorithm. For the incoming links, we used the topic title as a query to a search engine and took the top ranked results.

To place BEPs, we tried two different approaches: one was to place the BEP at the location of the anchor phrase in the target document (where the entire phrase, or part of it, appears); the other was to set the BEP to the beginning of the text block which contains terms similar to those of the text surrounding the anchor in the source document.

2 Link-The-Wiki Track

Regardless of the task the procedure for automated link discovery is the same and includes the following steps: identifying anchors, recommending a group of outgoing and incoming links, and locating BEPs for these links.

2.1 Anchor Identification

Anchor identification is the first step in link discovery. Identifying anchors can be done using two methods: best match; and partial match.

After an anchor is identified, a link, *a→d*, can be created. In this case *a* is an anchor and *d* is a corresponding target document.

2.2 Link Recommendation

2.2.1 Outgoing Links

Good algorithms for recommending outgoing links were seen at the first INEX Link-the-Wiki track in 2007 [1]. This year we employed the top two ranking algorithms from that (and subsequent) Link-the-Wiki tracks: Itakura's link mining (ICLM) algorithm [2] and Geva's page name matching (GPNM) algorithm [3]. The INEX 2009 document collection is, however, a much larger collection than the previous collection.

The ICLM algorithm relies on the pre-existing link graph in the Wikipedia. The anchor text (a) target document (d) pairs are all extracted from the collection. The document frequency of each anchor text is then computed. The algorithm proceeds by finding all anchor texts that exist within the orphaned topic document and ranking those on an anchor weight, γ:

$$\gamma = \frac{number\ of\ pages\ that\ has\ link(a \to d)}{number\ of\ pages\ that\ has\ text\ of\ anchor(a)} \qquad (1)$$

The GPNM algorithm generates a table containing the title and the id of each document in the collection. A sliding window with size varying from 1 to 12 terms is then run over the orphan topic document looking for titles from the table. These are ranked on target document title length; longer anchors having higher scores.

In an attempt to improve the performance of both algorithms we combined them into one. First, two sorted lists of outgoing links are created separately using the two algorithms. Links are assigned a score using following method:

$$Score(L) = Score_s(L) + Score_k(L) \qquad (2)$$

Where Score(L) is the score for link L, $\text{Score}_S(L)$ is the score from the GPNM algorithm and $\text{Score}_K(L)$ is a normalized ICLM γ:

$$Score_k(L) = \frac{\max(\gamma)-\min(\gamma)}{N} \times m \quad (3)$$

Where max(γ) is the highest γ value of any link in the GPNM list; min(γ) is the lowest γ value of any link in the list; N is length of the longest anchor in the list, and m is the number of terms in L.

Finally, links are ranked on Score(L). Either $Score_S(L)$ or $Score_K(L)$ might be zero if the anchor link only appears in one of the lists. Scores for links in both lists are boosted. The first 250 (for F2F) or 50 (for A2B) links of the list are selected as the links to return.

2.2.2 Incoming Links

Finding incoming links for a topic document is performed by retrieving the first 250 pages returned from a BM25 search engine. The orphaned document's title was used as the query terms.

2.3 BEP Location

Best entry points (BEPs) play an important role in providing readers direct access to relevant document passages [4]. However, deciding where the BEP should be is a difficult focused retrieval problem.

Our first approach to find the BEP is to find the first location of the anchor terms. There are two scenarios:

- If there is an exact match for the anchor in the destination page, the BEP is the offset of the first occurrence.
- If there is no exact match, then we use the location of the first term from the anchor text. If that cannot be found then we move on to the second term, and so in until a term is found. If no term is found then the start of the document is used.

Our second approach was a technique similar to that used in image matching. Given two images, the more features in those images that match, the more certain we can be that the two images depict similar objects. BEP finding can be treated as a feature finding problem. First, a text window of length 200 characters surrounding the anchor in the source document is used for creating a source text template. Terms are identified and Porter stemmed. These terms are the features. Next, a sliding window of the same length is passed over the target document, and features are extracted similarly. A score is calculated for the window by counting the number of matching features (stemmed terms). The window is moved forward 100 characters at a time and the score calculation for matched features is repeated. The beginning of the text block with the highest score is chosen as the BEP.

3 Link-the-Wiki Experiments

3.1 Link-the-Wiki Runs

Table 1. Link-the-Wiki runs.

Run name
QUT_LTW_F2F_SEA_BASELINE01
QUT_LTW_F2F_SEA_BASELINE02
QUT_LTW_F2F_SEA_BASELINE03 (unofficial)
QUT_LTW_F2F_SEA_01(disqualified)
QUT_LTW_F2F_SEA_02
QUT_LTW_F2F_SEA_04 (unofficial)
QUT_LTW_F2FonA2B_SEA_03(unofficial)
QUT_LTW_A2B_SEA_BASELINE01
QUT_LTW_A2B_SEA_BASELINE02
QUT_LTW_A2B_SEA_01
QUT_LTW_A2B_SEA_02

We submitted runs in this task at both levels: file-to-file (F2F) and anchor-to-bep (A2B). All the baseline runs (with BASELINE0X suffix) were created using the ICLM algorithm; and the other runs (with SEA_0X suffix) were generated using our new algorithm that combines ICLM and The GPNM algorithms.

The link table from ICLM algorithm included links in all pages including the orphan document (before it was orphaned) and so the link information in the orphan was removed from the table when calculating γ scores. In order to determine the impact of this necessary correction we submitted runs without the correction ("01" suffix). These runs are “cheating”. All other runs are correctly orphaned.

Table 2. Link-Te-Ara runs.

Run name
QUT_LTAra_A2B_SEA_BASELINE01
QUT_LTAra_A2B_SEA_BASELINE02

The baseline runs for Link-Te-Ara task were generated using the GPNM algorithm. The Te Ara document format is very different from that of the Wikipedia corpus. In a Te Ara page, there is no unique tag for the page title. For example, there may be only one <Name> tag, or many <*Name> tags.

A name to document pairing table for these name tags was created. The difference between Link-Te-Ara BASELINE01 and BASELINE02 lies in BEP identification: BASELINE01 uses the term matching technique for BEP identification; while BASELINE02 uses the text template matching technique.

Table 3. Link-Te-Ara-to-Wiki runs.

Run name
QUT_LTAraTW_A2B_SEA_BASELINE01

We only submitted only one run for Link-Te-Ara-to-Wiki task. This run was created using ICLM algorithm.

Separately, we submitted two further unsuccessful runs. The first was the ICLM algorithms: OTAGO_LINKPROBABILITY_A2B. The second was a modified ICLM that was generated by taking a proxy log of university of Otago student Wikipedia use and augmenting γ with a weight based on the number of times the link was clicked by a user: OTAGO_LINKPROBABILITYANDCLICKRATE_V1_A2B.

In this second run the new γ (η) was computed thus:

$$\eta = \gamma \times \frac{number\ of\ anchor\ text\ clicks}{number\ page(with\ the\ anchor\ text) views}$$

Unfortunately this second experiment was unsuccessful due to implementation issues. We will further this line of investigation in future work.

3.2 Link-the-Wiki Results

Since there are (at time of writing) no manual assessments or ground-true for the Link-Te-Ara task and the Link-Te-Ara-to-Wiki task, only the results from Link-the-Wiki task are discussed in this section.

The evaluation results for outgoing links on the file-to-file and anchor-to-bep topics are presented in figures 1 and 2 respectively. The results for incoming links are presented in figures 3 and 4. The results shown in these four figures are against the automatic assessments (links in the topic documents before orphaning). The results of our unofficial runs are included for comparison.

Among our runs which are correctly orphaned the best ones, marked as black curves shown in all the plots, indicate encouraging performance of our system. Figure 1 demonstrates that our run QUT_LTW_F2F_SEA_BASELINE02 has the highest score (for the correctly orphaned topic) run. However, run QUT_LTW_F2F_SEA_02 using the new algorithm has lower accuracy than the baseline runs have, even though still higher than others.

In figure 2 our system is out-performed by Waterloo's run in submissions for 50 outgoing links in file-to-file level on 33 topics for anchor-to-bep task.

Figure 5 presents the precision-recall curves for the unofficial runs of three different systems. The curve with the highest accuracy is from the Otago's system with the implementation using the ICLM algorithm. This figure indicates there might be a faulty in implementing ICLM algorithm in our system, since our run QUT_LTW_F2FonA2B_SEA_03 achieves lowest accuracy in terms of precision and recall comparing with others.

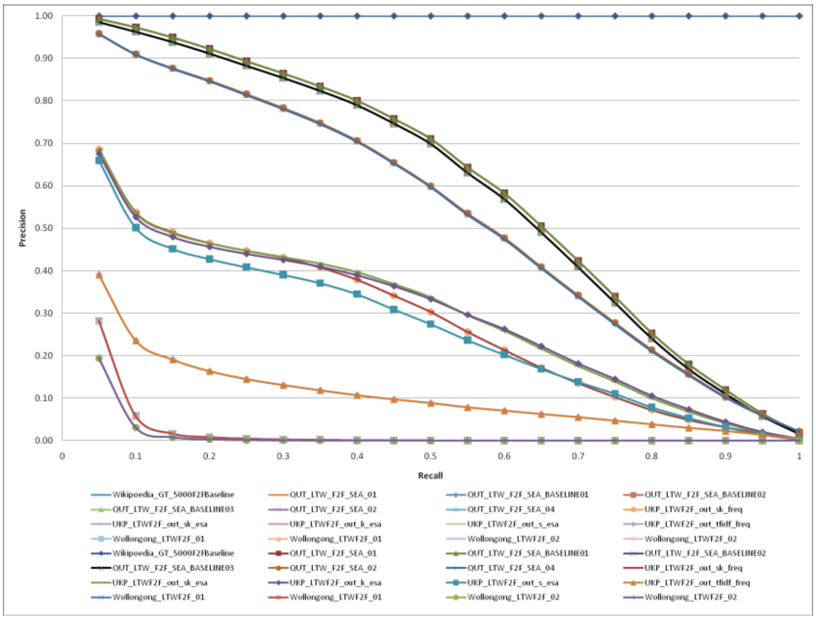

Fig. 1. Link-the-Wiki automatic outgoing F2F assessment on 5000 F2F topics.

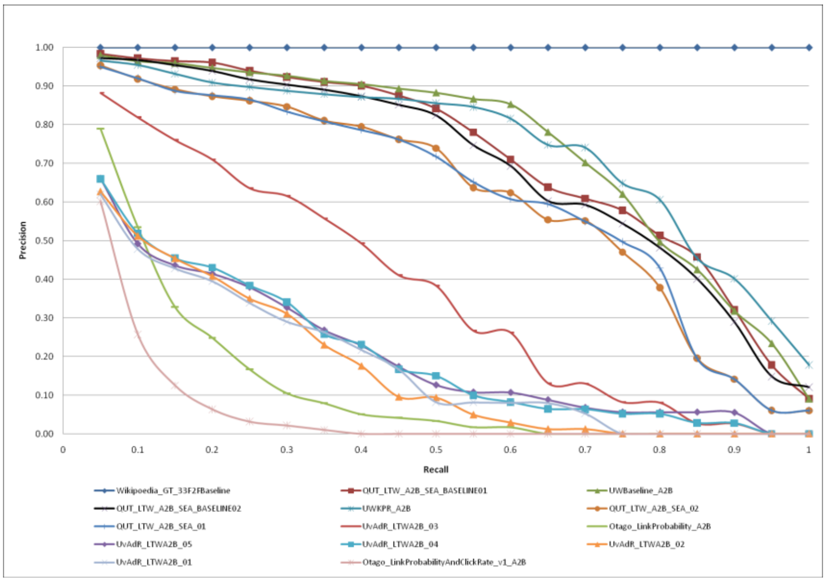

Fig. 2. Link-the-Wiki automatic outgoing F2F assessment on 33 A2B topics.

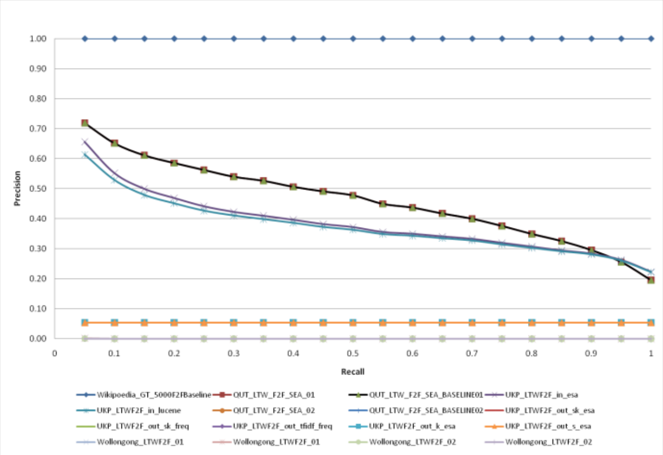

Fig. 3. Link-the-Wiki automatic incoming F2F assessment on 5000 F2F topics.

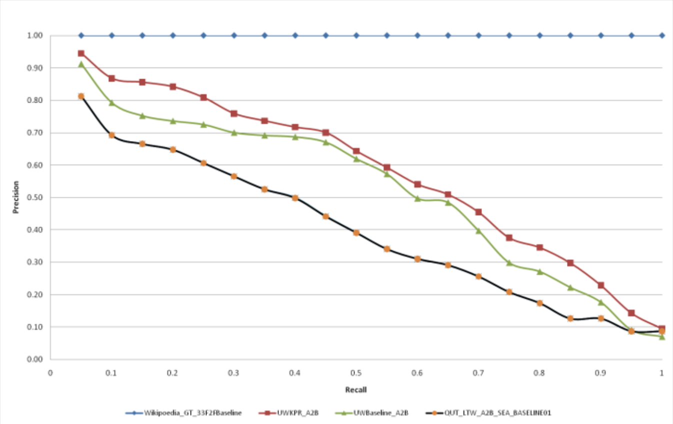

Fig. 4. Link-the-Wiki automatic incoming F2F assessment on 33 A2B topics.

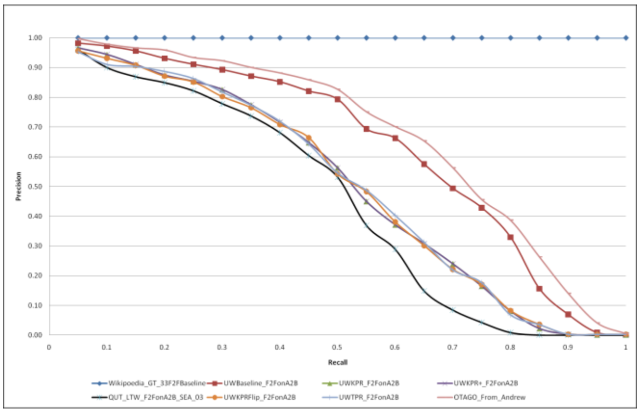

Fig. 5. Unofficial automatic outgoing F2F assessment on 33 A2B topics.

4 Conclusions and Future Work

In our link discovery system, we implemented the ICLM algorithm proposed in the first LTW track, and a new method that combines both ICLM and GPNM algorithms to generate outgoing links. Our results from the official evaluation of outgoing and incoming links show reasonable good performance of our system. Using traditional information retrieval technique on Wikipedia XML collection for creating incoming links is every effective.

The new hybrid method for recommending outgoing links doesn't work as well as the original ICLM algorithm. Finding out the reason for the degraded performance of the hybrid approach could be treated as our remaining task for next round of Link-the-Wiki evaluations in 2010.

References

1. Huang, D., Xu, Y., Trotman, A., Geva, S.: Overview of INEX 2007 Link the Wiki Track. Focused Access to XML Documents 373-387 (2008)
2. Itakura, K., Clarke, C.: University of Waterloo at INEX2007: Adhoc and Link-the-Wiki Tracks. Focused Access to XML Documents 417-425 (2008)
3. Geva, S.: GPX: Ad-Hoc Queries and Automated Link Discovery in the Wikipedia. Focused Access to XML Documents 404-416 (2008)
4. Reid, J., Lalmas, M., Finesilver, K., Hertzum, M.: Best entry points for structured document retrieval--Part I: Characteristics. Information Processing &amp; Management **42** 74-88 (2006)