Wikilinking and Wikisearching

Dylan Jenkinson, Andrew Trotman, Kai-Cheung Leung

Department of Computer Science

University of Otago

Dunedin

New Zealand

{dylan, andrew, kcleung}@cs.otago.ac.nz

**Abstract.** The University of Otago submitted six runs to the Link-the-Wiki track with the top run placing nth. Three element runs and three passage runs were submitted to the Relevant in Context task of the ad hoc track. The best Otago run was a whole-document run placing 7th. The best Otago passage run placed 13th while the best Otago element run placed 31st. There were a total of 40 runs submitted to the task. This result reinforced our prior belief that passages are better answers than elements and that the most important aspect of the focused retrieval is the identification of relevant documents..

Introduction

Otago participated in the Link-the-Wiki task in 2007 and produced runs that performed adequately however the results of Itakura & Clarke (Itakura and Clarke 2007) and of Geva (Geva 2007) were more successful at producing outgoing links. For 2008 Otago concentrated on reproducing and extending the work Itakura by including some of the findings of Geva, and including multiple targets per link.

To do this we first re-implemented their algorithm. This involved finding all phrases in the collection that were used as a link to another document, and recording this. We then found the most linked to document for each phrase, and stored that, along with a value representing the strength of the link. When a topic was processed, the 250 strongest links were used.

One problem with the results from this implementation was the occurrence of duplicate links between the orphan document and other documents in the collection. Such duplicates, where they pointed from the same anchor point in the orphan to the same anchor point in the target document, were considered as errors, and would have pulled the MAP down as a result. We removed these duplicates from our results. We also looked at weather paying attention to case would affect the results.

For incoming links, we tried several approaches. These included sorting the results based on the BM25 result from our search engine, learning weights for different terms in a search used to find documents related to the topic, and using the title of the topic for the search.

Stuff about how we did goes here.

Details of document collection used in the experiments.

Link Detection in the Wikipedia

The Link-the-Wiki task, first included in INEX in 2007, requires participants to automatically identify hypertext links between documents in the Wikipedia. The user model is that of a user who creates a new Wikipedia entry and would like to link that entry to pre-existing entries in the Wikipedia (and *vice versa*).

The production of a new article can be simulated by taking an existing Wikipedia document and removing all trace of it from the collection. Link identification software can then be run over the collection and the orphaned document. A comparison of the automatically generated links to the original collection gives some measure of the quality of the link detection system – that is, the original links are considered to be the gold-standard by which systems are compared.

Exactly this approach was taken in the INEX 2007 Link-the-Wiki track, and was used again for document-to-document linking in 2008. In 2008, 6600 documents (about 1% of the document collection) were randomly selected and orphaned for whole document link detection.

New in 2008 is the anchor-to-BEP linking task, in which the task is to identify the best orphan anchor from which to link from and the best-entry-point in the target document from which to link to. Unlike document-to-document linking, anchor-to-BEP linking requires manual assessment because the Wikipedia documents are not a priori marked-up in this way. For 2008, 50 anchor-to-BEP documents suggested by task participants and were orphaned for the experiment. A limit of 50 anchors per document was imposed (for practical reasons) and at most each anchor could link to 5 locations in the Wikipedia.

We examine the problem of link identification by first examining outgoing links (from the orphan to the collection) then incoming links (from the collection to the document).

Outgoing Links

Although the Otago runs in 2007 were adequate, those of Itakura and Clarke (Itakura and Clarke 2007) were substantially better – we chose, therefore, to investigate methods of improving their technique. It should be noted that the Itakura & Clarke algorithm relies on a pre-existing heavily interlinked document collection (such as the Wikipedia). In the case where no prior links exist in the collection the techniques of Geva (Geva 2007) which were also successful in INEX 2007 can be used.

The Itakura & Clarke Algorithm

The Itakura & Clarke algorithm relies entirely on pre-existing links between documents within the document collection. Of the link types available in the collection, only the <collectionlink> type is utilized because the other link types do not link between two documents in the collection (for example, a <wikipedialink> links from a document in the collection to a document in the Wikipedia that is not in the INEX collection.

Initially a list of all the links within the document collection is created. This is generated by parsing each document in the collection and extracting the anchor text of the link and the target document id.

Next and from the output of the previous stage, a list of target documents is created for each unique anchor text in the collection. For a given anchor text in the collection, the most frequent target is most likely to be the correct target.

For each anchor text / target pair a strength value ( is constructed

*=np/af*

where *np* is the number of documents that link from the anchor to the target and *af* is the number of pages in which the anchor text occurs.

An orphaned document is then parsed and the first location of each anchor in the pre-generated list is located. For overlapping anchors (for example, “Lennon” and “John Lennon”) the longest possible anchor is chosen as a longer anchor is more likely to be correct than a short anchor. A limit of 250 anchors per document was enforced by the Link-the-Wiki track definition.

Small Improvements

After implementing the Itakura & Clarke algorithm verbatim we identified a number of small improvements.

The algorithm defines the anchor text as all text occurring between the tags regardless of case and punctuation. Anchor texts often contain punctuation at the end thus creating a distinction between “John Lennon” and “John Lennon.”. We stripped punctuation from the anchors.

Anchor texts beginning at the start of a sentence are capitalized for grammatical reasons so the algorithm converts the text into lower case. Unfortunately this results in a distinction between “unfinished music” and “Unfinished Music” (the two part experimental work by John Lennon). Geva (Geva 2007) identifies the important of case in link detection so we dropped the case conversion step.

Finally, we over-weighted  for terms containing capitalization by adding a constant,  where terms in the orphan were found capitalized. We expect those terms to be proper nouns and thus Wikipedia entries.

Figure 1 compares our implementation of the algorithm to the original on the INEX 2007 Link-the-Wiki topics. The line labeled “Waterloo” is the Itakura & Clarke run as submitted. Removing punctuation (Alphanumeric) from the anchor list improves the algorithm, removing case folding (Case Sensitive) leads to further improvements. Our best run (Weighed) included punctuation removal, case sensitivity, and weighted 

Figure 2 shows the effect of  on precision, a value of 0.3 would have been best for early precision, but a value of 0.1 hold the precision longer resulting in the highest mean average precision.



**Fig. 1.** **Small improvements on the Itakura & Clarke algoritm (Waterloo) are seen when punctuation is removed (Alphanumeric), when case folding is removed (Case Sensitive) and when uppercase anchors are preferred over lowercase anchors (Weighted).**



**Fig. 2. Effect of varying  on the precision. Small value of  (0.3) is best for early precision but a very small score (0.1) holds the precision higher longer (best for MAP).**

Best Entry Points

Several studies have shown the best entry point for Wikipedia documents is the start of the document. (Jenkinson and Trotman 2007; Kamps, Koolen et al. 2007). We use this prior result and target start of the document as the link best entry point.

Multiple Targets

The Link-the-Wiki task specification for 2008 allowed at most 5 targets for each anchor point. The Itakura & Clarke algorithm was, consequently, extended to so that the  value was computed for not just the most common target, but also for all targets of an anchor text. The  values represent the probability of the target document being the correct target; For our runs we simply targeted the top 5  for each anchor text.

Incoming Links

The best Otago run at INEX 2007 achieved an excellent early precision (P@5) score of 0.751. Our approach for 2008 was to extend and improve on this method.

The Otago 2007 Algorithm

The algorithm for detecting incoming links relies on the belief that links should be reciprocal – that is, in there should be a link from document A to document B then there should be a link from B to A. If this is the case the discovering incoming links is the reciprocal operation of discovering outgoing links.

For each unique term (excluding stop words) in the orphaned document the Otago 2007 algorithm (Jenkinson and Trotman 2008) computes the actual frequency of that term, *af*

*af = tf / dl*

where *tf* is the number of occurrences of the term in the orphan and *dl* is the length of the orphan (in terms); and the expected frequency, *ef*

*ef = cf / (df \* ml)*

where *cf* is the number of occurrences of the term in the collection, *df* is the number of documents containing the term and *ml* is the mean length of a document. Ranking the terms in the orphan by ratio of *af* to *ef*

*st = af / ef*

provides a list of terms in order of occurrence relative to expected occurrence. If this ratio is larger than one the term occurs in the document more often than expected, if it is less than one it occurs less frequently than expected. The top ranked terms are representative themes of the document and are used to construct queries of the collection. The results of these queries are documents relevant to the themes of the orphan and therefore the two should be linked (in both directions).

Improvements – Multiple Searches

For INEX 2007 queries were constructed by taking the top *n* terms from the list and performing a query, extracting the top *n* \* 50 results and then concatenating to that list the top n\*50 results until a total of 250 results were found. That is, for *n*=2, three searches were performed, the first identifying the top 100 results and the second identifying the next 100 results, and the last identifying the remaining 50 results. There was no theoretic justification for this approach; it was motivated by time constraints. It is of note, however, that it was not an unsuccessful approach.

By merging the results of each separate query on the rsv (in this base BM25), good targets that match other than the top theme will be placed high in the results list. This approach, might, also place documents that are good matches for non-key themes high in the results list because of a high rsv with respect to a non-key term.

To alleviate this problem the BM25 score for each search term can be weighted by a weight for each key term in the query. The strength of a term with respect to the orphan has already been computed (*st*) and so that value was used in our experiments.

Our best run at INEX 2007 used two searches of 4 term each producing a total of 250 results in the results list. We experimented with the number of search terms using the 2007 algorithm, merging, and weighted merging and the best number was 2.

The results are shown in Table 1. The best runs submitted to INEX 2007 (by any participant) achieved a score of 0.484 and is listed for comparative purposes. Our best run at INEX 2007 achieved a score of 0.339 which is better than the score achieved by result merging, 0.319, but not as good as the 0.350 achieved by weighted result merging. Figure 3 shows the early precision scores for the same three techniques. Of particular interest is that although the MAP score for weighted merging is highest, the early precision scores of the Otago 2007 run are highest.

**Table 1. MAP scores for different approaches to multiple searches. The weighted merging of queries containing 2 terms each achieved a better score than the best Otago 2007 run, however not as good as the best run submitted by any institute.**

|  |  |
| --- | --- |
| **Run** | **MAP** |
| Top INEX 2007 run | 0.484 |
| Weighted merge | 0.350 |
| Otago 2007 | 0.339 |
| Merged | 0.319 |



**Fig. 3. Early precision scores for the three merging techniques. Although the MAP of weighted merge is highest, the early precision of Otago 2007 is highest.**

Improvements – Single Searches

With the multiple search technique the contribution of each separate search to the final precision score is unclear. It is also unclear whether or not a better approach is to simply perform one search with the given number of terms and to use the top 250 results.

Two experiments were conducted: in the first, *n* search terms were used and *n* \* 50 results were retrieved; in the second, *n* search terms were used but the full 250 results were retrieved. The results were compared the multiple search technique without merging and without weighting.

Figure 4 shows the contribution of the first search is a substantial proportion of the final result of the multiple search approach. It also shows improvements on the multiple search technique when the full 250 results are retrieved. The improvements decrease as the number of terms per query increases to 5 as the number of documents retrieved per query in the multiple query approach tends to the full 250.



**Fig. 4. A comparison of the multiple search technique to the single search technique suggests that the single search technique is best.**

Weighted Search Terms

The experiments examining multiple searches showed that MAP could be improved if the search terms were weighted by *st*. Improvements are therefore expected in the single search approach if the individual search terms in a single query are weighted. The weights could be taken from the *st* score, but we chose to learn weights using Genetic Algorithms {Holland, 1975 #245}.

Trotman {Trotman, 2005 #486} and later Robertson et al. {Robertson, 2004 #596} modify the term frequency component of BM25 to include a separate weight for each structure within a document. We use their approach to weight term frequencies based not on the structure, but on the position of the term in the query (where query terms are sorted in decreasing *st* score). The new term frequency score use in the BM25 equation, *tf*, is given by

*tf = tft \* cq*

where *tft* is the true term frequency of the term in the document; and *cq* is a constant weight for a term at position *q* in the query, varying from 0 to 1.

If the weight of *cq* is 0 then the search term will be discarded from the query. If it is 1 then the true term frequency will be used, otherwise the influence of the term frequency will be scaled by *cq*. Good values for *cq* are expected to decrease as a function of distance from the start of the query, eventually reaching 0 when adding new terms begins to ambiguate the query.

We experimented with learning weights for queries of length between 2 and 10 search terms[[1]](#footnote-2). The population size was 50, crossover rate was 0.9, mutation rate was 0.05, and reproduction rate was 0.05. The learning was run for 10 generation. Elitism was used. Several iterations of the learning were conducted and the best weights of the best run were recorded.

For the best MAP score achieved for queries ranging from 2 to 10 search terms, Table 2 shows the weights that were learned. It can be seen from this table that the first two terms are responsible for the majority of the performance.

Figure 5 shows that weighting search terms results in an increase in precision for all cases we tried with the exception of a single search term. It should be noted that in the experiments over-fit the weights to the orphan documents, but there is an insufficient number of orphans to conduct a traditional learn / validate / evaluate experiment.

**Table 2.** Best learned weights for different queriy lengths

|  |  |
| --- | --- |
| **Search Terms** | **Weights (from first to last term)** |
| 2 | 0.96, 0.95 |
| 3 | 0.99, 0.96, 0.04 |
| 4 | 0.97, 0.73, 0.05, 0.06 |
| 5 | 0.95, 0.83, 0.14, 0.1, 0.01 |
| 6 | 0.89, 0.97, 0.44, 0.41, 0, 0.06 |
| 7 | 0.8, 0.95, 0.75, 0.29, 0, 0.07, 0.25 |
| 8 | 1, 0.88, 0.14, 0.05, 0, 0.22, 0.08, 0.19 |
| 9 | 0.87, 0.81, 0.36, 0.26, 0, 0.22, 0.29, 0.2, 0.01 |
| 10 | 0.9, 0.99, 0.77, 0.55, 0.35, 0.08, 0.19, 0.16, 0, 0.19 |



**Fig. 5. Effect of weighting individual search terms in the query.**

Other Sources of Search Terms

The experiments thus far suggest that the best approach is to perform a single search using a small number (two or three) highly representative search terms to identify document that should point to the orphan. The approach to identifying terms involved identifying document themes by simply text processing techniques. Wikipedia documents, however, are structured and include a title as well as a brief overview of the content of the document. These document structures might be used as a method of identifying good representative document-thematic terms. The whole document has been used by others {Fachry, 2007 #1005}.

The title of the Wikipedia document is held between <name> tags. These were processed to remove duplicate search terms and stop words, then used as relevance ranked queries.

The overview of the Wikipedia document occurs as an untitled section before the first titled section. It was extracted by using all text before the first <title> tag of the document, or if there was no <title> tag then the whole document was used.

The full-text of the Wikipedia document can easily by extracted by removing all XML tags from the document.

Figure 6 shows the effect on early recall of the different techniques. Selecting terms from the whole document is better than using the title which is better than the overview which in turn is better than the whole document. However, the result is somewhat different when the MAP scores are compared; Table 3 presents the MAP scores and it can be seen that using the title is better overall than the other approaches, even bettering the weighted merge approach from above.



**Fig. 6.** Different sources of search terms. The title is a more effective source of terms than the overview which is better than the whole document. For early precision the best source was the approach used by Otago at INEX 2007

**Table 3.** MAP scores of the runs using terms from different parts of the document

|  |  |
| --- | --- |
| **Run** | **MAP** |
| Title | 0.410 |
| Overview | 0.143 |
| Document | 0.080 |
| Otago 2007 | 0.339 |
| Weighted merge | 0.350 |

Otago Link-the-Wiki 2008 Runs

Problematic and systemic with our experiments is the tradeoff of early precision with mean average precision. The best methods to choose is dependent on the metric being used to score the runs. MAP was the metric used in 2007 and we assumed its use for 2008.

File-to-file linking

Three runs were submitted to the file-to-file linking task:

capConstant-SingleSearchWeighted: outgoing links were identified using the Otago version of Itakura & Clarke with  = 0.1. Incoming links were identified using the weighted merge method with 4 search terms and weights of 0.97, 0.73. 0.05 & 0.06.

capConstant-TitleOnly: outgoing links were identified using the Otago version of Itakura & Clarke with  = 0.1. Incoming links we were identified using the title of the orphan.

nonCap-FirstPara:

The third run uses the nonCapConstant method, which is Kelly's but with

the deduplication and such added in, and for incoming uses the whole

first paragraph of the topic in a search to find the appropriate files.

Anchor-to-BEP linking

capConstant-SingleSearch-A2B

capConstant-TitleOly-A2B

nCapConstant-WholeDocument-A2B

. The first

two runs here are the same as the first two runs in file to file, with

the same weights and constants used. The third run uses the

nonCapConstant method, the same as the third file to file run, and uses

a the whole topic document as the search for incoming links.

Focused Searching in the Wikipedia

Passages

The approach taken by Otago at INEX 2007 [[CITE]] was two step. First, relevant documents were identified using BM25. Second, all the occurrences of all the search terms with a document were identified (stemming with Porter’s algorithm) and a fixed sized window of 300 characters placed on the centroid. The window was 300 words in length. The centroid was defined as the mean of the term locations within the document, or alternatively the mean of those within one standard deviation of the true mean.

Huang et al. {Huang, 2006 #980} examined techniques for identifying relevant passages within a relevant document and converting those into elements by taking the smallest element that fully enclosed the passage. The Kullback-Leibler model was the most effective method they used:

where *W* is window within *D* (the document) and *t* is a search term of *Q* (the query) and

and

Where *tfD* is the number of occurrences of *t* in *D* and |*D*| is the length of document *D* (and likewise for *tfW* with respect to the window, *W*).

Several strategies for choosing the window were examined. The sliding non-overlapping window of size 400 words was shown to be effect on the INEX IEEE document collection (measured with MAep and iMAep).

Itakura and Clarke {Itakura, 2007 #966} suggest that methods of identifying elements from passages are not as effective as methods of identifying elements directly. This is, in part, because conversion from a passage to an element usually involves increasing the size of the passage and this extra text is expected to be non-relevant (by the passage retrieval algorithm). That is, the conversion from a passage to an element is unlikely to affect recall but is likely to decrease precision. If this is the case then the prior reported result of Huang et al. is understated.

Elements

Beigbeder [[CITE]] proposes a method of scoring elements based on fuzzy proximity. If a document contains one occurrence of one search term, then the fuzzy proximity (*fp*) to term occurrence, t, at location *p* is

If the document contains more than one term occurrences of the same term then the fuzzy proximity is defined as the fuzzy proximity to the closest term occurrence. If the document contains multiple search terms then the fuzzy proximity is defined as the minimum fuzzy proximity to any search term.

Documents

At INEX 2007 RMIT University ad hoc submission demonstrated that a full-document run could be more effective at focused retrieval than a focused run {Fuhr, 2008 #1006}.

Otago ad hoc 2008 Runs

Three runs were submitted to the Relevance-in-Context passage task:

WHOLEDOC\_PASSAGE: Documents were ranked using BM25 and the whole document was returned as the passage.

DYLAN\_200: Documents were ranked using BM25 then a fixed sized window of 200 words was placed on the centroid of the search terms within the document. The standard deviation method was used to compute the centroid.

SW\_KL\_200: Documents were ranked using BM25 and the Kullback-Leibler method with sliding window was used to identify a relevant passage.

Three runs were submitted to the Relevance-in-Context element task:

WHOLEDOC: Documents were ranked using BM25 and the whole document was returned as an element (this run is identical to WHOLEDOC\_PASSAGE, but was submitted as a sanity check).

BEIGBEDER\_ORIG: Documents were ranked using BM25 and then elements were selected using a faithful implementation of Beigbeder’s algorithm.

BEIGBEDER\_IDF: Documents were ranked using BM25 and then elements were selected using the IDF weighed version of Beigbeder’s algorithm.

The results are presented in Table XXYY where it can be seen that WHOLEDOC and WHOLEDOC\_PASSAGE do, indeed, score the same thus passing the sanity check. The passage algorithms are superior to the element algorithms with the Kullback-Leibler approach bettering the Otago 2007 approach by a very small amount. The IDF enhancement to Beigbeder’s algorithm increases the precision substantially, but not sufficiently to better the passage runs.

Table XXYY: ad hoc results

|  |  |  |
| --- | --- | --- |
| **Run** | **Type** | **MAgP** |
| WHOLEDOC\_PASSAGE | Passage | 0.192 |
| WHOLEDOC | Element | 0.192 |
| SW\_KL\_200 | Passage | 0.183 |
| DYLAN\_200 | Passage | 0.182 |
| BEIGBEDER\_IDF | Element | 0.149 |
| BEIGBEDER\_ORIG | Element | 0.107 |

Conclusions

Some conclusions go here.

Acknowledgements

Funded in part by a University of Otago Research Grant.

References

Geva, S. (2007). GPX: Ad-Hoc Queries and Automated Link Discovery in the Wikipedia. Focused Access to XML Documents: 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2007 Dagstuhl Castle, Germany, December 17-19, 2007. Selected Papers, Springer-Verlag: 404-416.

Itakura, K. Y. and C. L. Clarke (2007). University of Waterloo at INEX2007: Adhoc and Link-the-Wiki Tracks. Focused Access to XML Documents: 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2007 Dagstuhl Castle, Germany, December 17-19, 2007. Selected Papers, Springer-Verlag: 417-425.

Jenkinson, D. and A. Trotman (2007). Wikipedia Ad Hoc Passage Retrieval and Wikipedia Document Linking. Focused Access to XML Documents: 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2007 Dagstuhl Castle, Germany, December 17-19, 2007. Selected Papers, Springer-Verlag: 426-439.

Jenkinson, D. and A. Trotman (2008). Wikipedia Ad Hoc Passage Retrieval and Wikipedia Document Linking. Focused Access to XML Documents: 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2007 Dagstuhl Castle, Germany, December 17-19, 2007. Selected Papers, Springer-Verlag: 426-439.

Kamps, J., M. Koolen, et al. (2007). Where to Start Reading a Textual XML Document? 30th ACM SIGIR Conference on Information Retrieval, Amsterdam, The Netherlands.

1. In the case of a single search term the weight has scaling effect which does not affect the relative rank order of the results; and so has no effect on MAP. [↑](#footnote-ref-2)