Wikisearching and Wikilinking

Dylan Jenkinson, Kai-Cheung Leung, Andrew Trotman

Department of Computer Science
University of Otago
Dunedin
New Zealand

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

Abstract. The University of Otago submitted three element runs and three passage runs to the Relevance-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. The ad hoc 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. Six runs were submitted to the Link-the-Wiki track. At time of writing the results had not been published.

1. Introduction

Otago participated in the Relevance-in-Context task of the ad hoc track submitting six runs, three passage and three element runs. The passage runs compared the Otago 2007 algorithm to a previous algorithm examined by Otago, the Kullback-Leibler model, and to whole document retrieval. The result suggest that whole document is better than passage retrieval and that there is little difference between the other two algorithms.

Otago also participated in the Link-the-Wiki track, preferring a variant of the Itakura & Clarke algorithm for outgoing links, and searching for the orphan title for documents that should link to the orphan. At the time of writing the results for the track had not been published.

2. Wikisearching

2.1. Passages

2.1.1. The Otago 2007 Algorithm

The approach taken by Otago at INEX 2007 [1] 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 words placed on the centroid. 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.

2.1.2. The Kullback-Leibler Algorithm

In earlier experiments at Otago, Huang et al. [2] 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. Of the passage selection methods examined, the Kullback-Leibler model was the most effective:

$$\mathrm{KL}(\mathrm{W}|\mathrm{Q}) = \sum_{\mathrm{t}\in\mathrm{Q}} \mathrm{P}(\mathrm{t}|\mathrm{W})\log\left(\frac{\mathrm{p}(\mathrm{t}|\mathrm{W})}{\mathrm{p}(\mathrm{t}|\mathrm{D})}\right)$$

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

$$p(\mathrm{t}|W) = \frac{tf_W + 0.5}{|\mathrm{W}| + 1}$$

and

$$p(\mathrm{t}|D) = \frac{tf_D + 0.5}{|\mathrm{D}| + 1}$$

where tf_D is the number of occurrences of *t* in *D* and |*D*| is the length of document *D* (and likewise for tf_W 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 effective on the INEX IEEE document collection (measured with MAep and iMAep).

Itakura and Clarke [3] suggest that methods of identifying elements from passages are not as effective as methods of identifying elements directly. This is, in part, because the 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 and a comparison of Kullback-Leibler to Otago 2007 is necessary to progress our work.

2.2. Elements

2.2.1. The Beigbeder Algorithm

Beigbeder [4] 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*, for location *p* is

$$fp = max\left(\frac{k-(p-t)}{k}, 0\right)$$

If the document contains more than one term occurrence of the same term then the fuzzy proximity is defined as the fuzzy proximity to the closest term occurrence (that is, *max*(*fp*) with respect to that term). If the document contains multiple search terms then the fuzzy proximity is defined as the minimum fuzzy proximity to all search terms.

The fuzzy score of an element in a document is computed as the sum of fuzzy proximity scores for each term in the element, normalized by the length of the element. However, as the documents are hierarchically structured, if a search term occurs in the title of a section then the fuzzy proximity of a term in the element to the search term in the title is defined as 1.

2.2.2. Small Improvements

Beigbeder's algorithm treats all terms as equal whereas it is usual for scoring algorithms to weight terms differently. The algorithm is thus extended to include some aspect of the strength of a search term (IDF). The IDF weighted fuzzy proximity, *fp'* is given by

$$fp' = IDF * max\left(\frac{k-(p-t)}{k}, 0\right)$$

the variant of IDF chosen is

$$IDF = \frac{N-n+0.5}{n+0.5}$$

where *N* is the number of documents in the collection and *n* is the number of documents in which the term occurs.

Problematically, if a search term is missing from the document then the fuzzy proximity to that term is always zero and so no part of the document is considered relevant (due to the *min*() function). Using the sum of fuzzy proximity weights in place of the minimum overcomes this problem.

The Beigbeder algorithms is of general interest as it is a method of identifying relevant elements as a function of term proximity, and can be extended to identify relevant passages. A comparison of the original Beigbeder algorithm and the Otago variant; as well as to the Otago passage runs will help answer the question of whether passages or elements are the best result to the Relevance-in-Context task.

2.3. Documents

At INEX 2007 an RMIT University ad hoc submission demonstrated that a full-document run could be more effective at focused retrieval than a focused run [5].

Geva and Winters[1] suggest this is because the F measure of recall and precision pre-selects choosing whole documents as 100% recall within a document can be easily realized. Whole document runs were, therefore, submitted for comparison to the focused retrieval runs.

2.4. Otago ad hoc 2008 Runs and Results

Three runs were submitted to the Relevance-in-Context passage task. In all cases documents were identified using BM25 (k_1=1.2, k_3=7.0, b=0.75) and then one passage was identified for each document in the top 1500 documents. The rank order of the final results was BM25. Stemming was not used.

WHOLEDOC_PASSAGE: The whole document was returned as the passage.

DYLAN_200: 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: The Kullback-Leibler method with a sliding window of 200 words was used to identify a relevant passage.

Three runs were submitted to the Relevance-in-Context element task, BM25 was used to identify the top 1500 documents, one element was identified, and the results re-ranked based on the Beigbeder score. For these experiments *k*=200.

WHOLEDOC: The whole document was returned as an element (this run is identical to WHOLEDOC_PASSAGE and was submitted as a sanity check).

BEIGBEDER_ORIG: Elements were scored using Beigbeder's algorithm.

BEIGBEDER_IDF: Elements were scored using the IDF weighed version of Beigbeder's algorithm. Due to a bug in our code we actually implemented the product of the sum of the IDF and *fp* scores in place of the sum of the product.

2.5. Wikisearching Results

The results are presented in Table 1 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.

[1] Private communications

Table 1. Ad hoc Relevance-in-Contest task 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

3.0 Wikilinking

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 applied to 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 document-to-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 (BEP) 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 typically not *a priori* marked-up in this way. For 2008, 50 anchor-to-BEP documents were 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.

Two separate problems exist with identifying links, the identification of outgoing links (from the orphan to the collection) and the identification of incoming links (from the collection to the document).

3.1. Outgoing Links

Although the Otago runs in 2007 were adequate, those of Itakura & Clarke [6] were substantially better – effort was, therefore, spent investigating 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 [7] which were also successful in INEX 2007 can be used.

3.1.1. 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 (γ) is constructed

$$\gamma = \frac{np}{af}$$

where *np* is the number of documents that link from the anchor to the target and *af* is the number of documents 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.

3.1.2. Small Improvements

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

The algorithm defines the anchor text as all text occurring between the tags, converted to lowercase, and including 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 thus conflating these two cases.

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 and Yoko Ono). Geva [7] identifies the importance of case in link detection so the case conversion step was dropped.

Finally, over-weighting γ for capitalized terms in the orphan will help identify proper noun conflicts (such as Unfinished Music). A capitalization constant, Π, is added to γ where terms in the orphan were found capitalized.

Figure 1 compares the improvements to the original algorithm using 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. Weighting (Weighed) includes punctuation removal, case sensitivity, and weighted γ, and was the best experimental run on the 2007 orphans.

Figure 2 shows the effect of Π on precision, a value of 0.3 is best for early precision, but a value of 0.1 holds the precision longer resulting in the highest mean average precision.

3.1.3. Best Entry Points

Several studies have shown the best entry point for Wikipedia documents is the start of the document. [1, 8]. No further investigation was performed on BEPs.

3.1.4. 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 so that the γ value was computed for not just the most common target, but also for all targets of an anchor text. The γ values represent the probability of the target document being the correct target; consequently choosing the top five documents (by γ) for each anchor text satisfies the track requirements.

3.2. Incoming Links

The best Otago run at INEX 2007 achieved an excellent early precision (P@5) score of 0.751. The experiments described in this section were conducted in an effort to improve the overall performance (MAP) and were conducted on the 2007 Link-the-Wiki oprhans.

3.2.1. The Otago 2007 Algorithm

The algorithm for detecting incoming links relies on a simple theme extraction technique used to identify the semantic content of the document.

For each unique term (excluding stop words) in the orphaned document the Otago 2007 algorithm [1] computes the actual frequency of that term, *af*

$$af = \frac{tf}{dl}$$

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

$$ef = \frac{cf}{df * ml}$$

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)*,

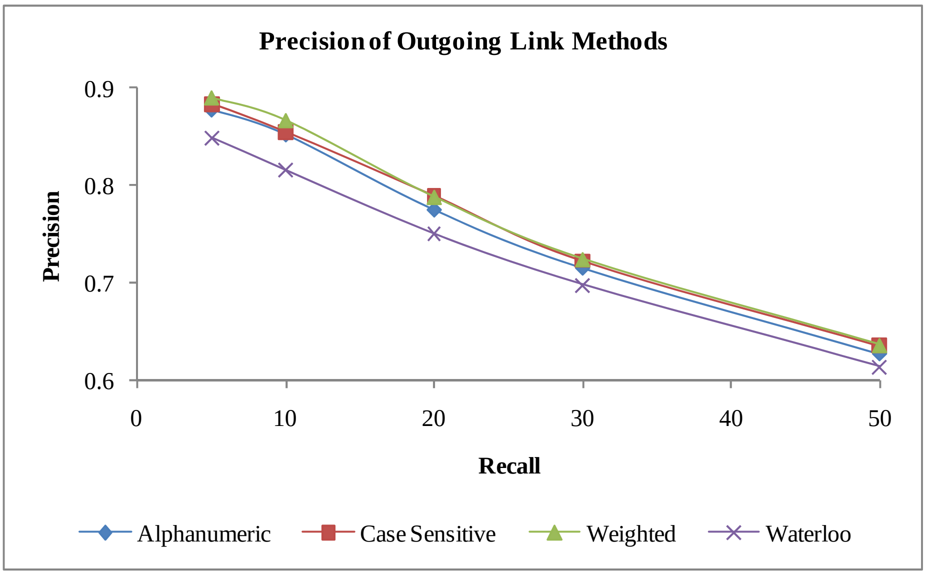

Fig. 1. Small improvements on the Itakura & Clarke algoritm (Waterloo) are seen when punctuation is removed (Alphanumeric), when case folding is removed (Case Sensitive) and when uppercase anchors are preferred over lowercase anchors (Weighted).

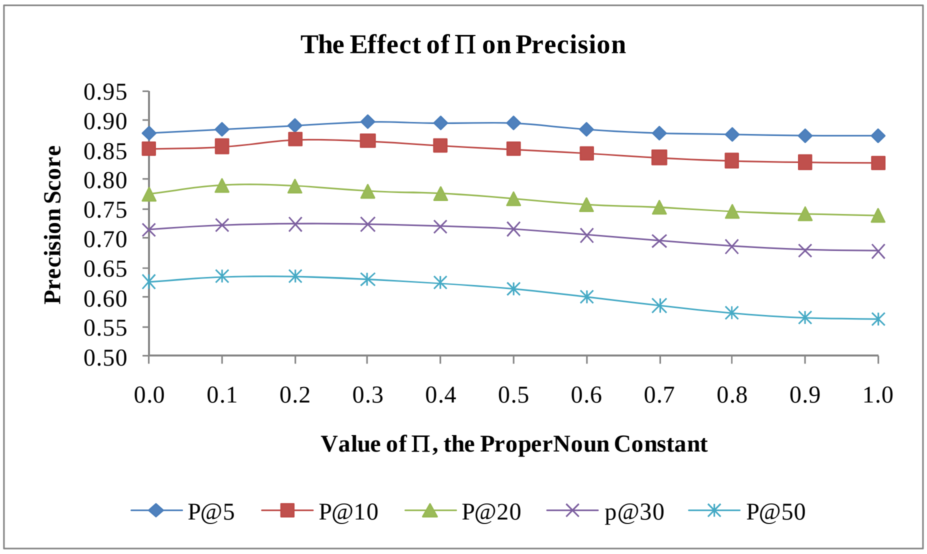

Fig. 2. Effect of varying Π on the precision. Small value of Π (0.3) is best for early precision but a very small score (0.1) holds the precision higher longer (best for MAP).

$$st = \frac{af}{ef}$$

provides a list of terms in order of occurrence relative to expected occurrence. If this ratio is larger than 1 the term occurs in the document more often than expected, if it is less than 1 it occurs less frequently than expected. The top ranked terms are representative themes of the document and are used to construct queries. The results of these queries are documents relevant to the themes of the orphan and therefore these documents should link to the orphan.

3.2.2. Improvements – Multiple Searches

For INEX 2007, queries were constructed by taking the top *n* terms from the *st*-ordered term-list and performing a query, extracting the top *n* * 50 results and then concatenating them to the list of 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 case the BM25 score) 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. The strength of a term with respect to the orphan has already been computed (*st*) and so that is a reasonable value to choose.

The best Otago run at INEX 2007 used two searches of 4 terms each, producing a total of 250 results in the results list. Using merging and weighted merging on the 2007 orphans the best number was 2.

The results are shown in Table 2. The best runs submitted to INEX 2007 (by any participant) achieved a score of 0.484 and is listed for comparative purposes. The best Otago 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 2. 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

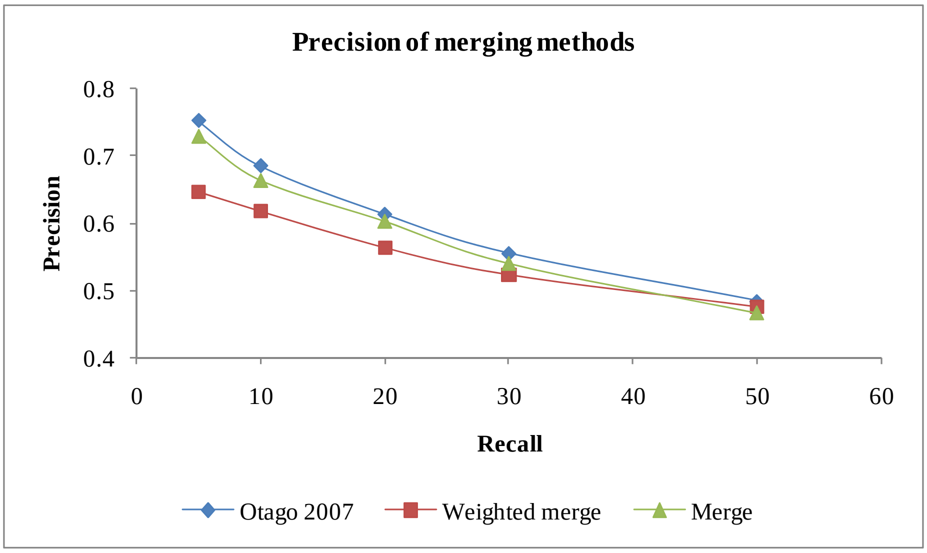

Fig. 3. Early precision scores for the three merging techniques. Although the MAP of weighted merge is highest, the early precision of Otago 2007 is highest.

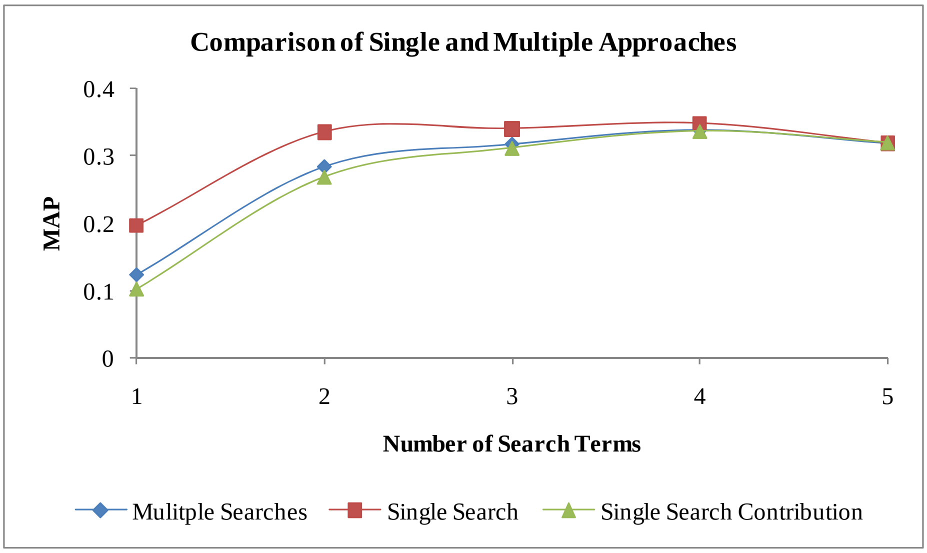

Fig. 4. A comparison of the multiple search technique to the single search technique suggests that the single search technique is best.

3.2.3. 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 first experiment computes the contribution of the first search to the multi-search whereas the second compares multi-searching with single-searching. The results were compared to 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 the superiority of the single 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.

3.2.4. 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 [9].

Trotman [10] and later Robertson et al. [11] 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

$$tf = tft * c_q$$

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

If the weight of c_q 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 linearly scaled by c_q. Good values for c_q are expected to decrease as a function of distance from the start of the query, reaching 0 when adding new terms creates an ambiguous query.

Experiments were conducted to learn weights for queries of lengths between 2 and 10 search terms[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. Many iterations of the learning were conducted and the best weights of the best run were recorded.

[2] In the case of a single search term the weight has a scaling effect which does not affect the relative rank order of the results; and so has no effect on MAP.

For the best MAP score achieved for queries ranging from 2 to 10 search terms, Table 3 shows the weights that were learned. It can be seen 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 tested cases (with the exception of a single search term). It should be noted that the experiments over-fit the weights to the orphan documents; unfortunately there is an insufficient number of orphans (in the 2007 set) to conduct a traditional learn / validate / evaluate experiment.

Table 3. 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

Table 4. 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

3.2.5. 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 simple 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, or the whole document (as seen by others [12]) might be used.

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

The full-text of the Wikipedia document can easily by extracted by removing all XML tags from the document, removing stop and duplicate words, and used as the query.

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 4 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.

3.3. Otago Link-the-Wiki 2008 Runs

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

3.3.1. File-to-file linking

Three runs were submitted, each used BM25 (k_1=0.421, k_3=242.61, b=0.498)

capConstant-SingleSearchWeighted: outgoing links were identified using the Otago version of Itakura & Clarke with $\Pi = 0.1$. 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 $\Pi = 0.1$. Incoming links we were identified using the title of the orphan.

nonCap-FirstPara: outgoing links were identified using the Otago version of Itakura & Clarke without Π. Incoming links were identified using the outline of the orphan.

3.3.2. Anchor-to-BEP linking

capConstant-SingleSearch-A2B: same as capConstant-SingleSearchWeighted.

capConstant-TitleOnly-A2B: same as capConstant-TitleOnly.

nCapConstant-WholeDocument-A2B: same as nonCap-FirstPara, but using the whole document for the query.

3.4 Wikilinking Results

At time of writing the results of the Link-the-Wiki track for 2008 had not been published.

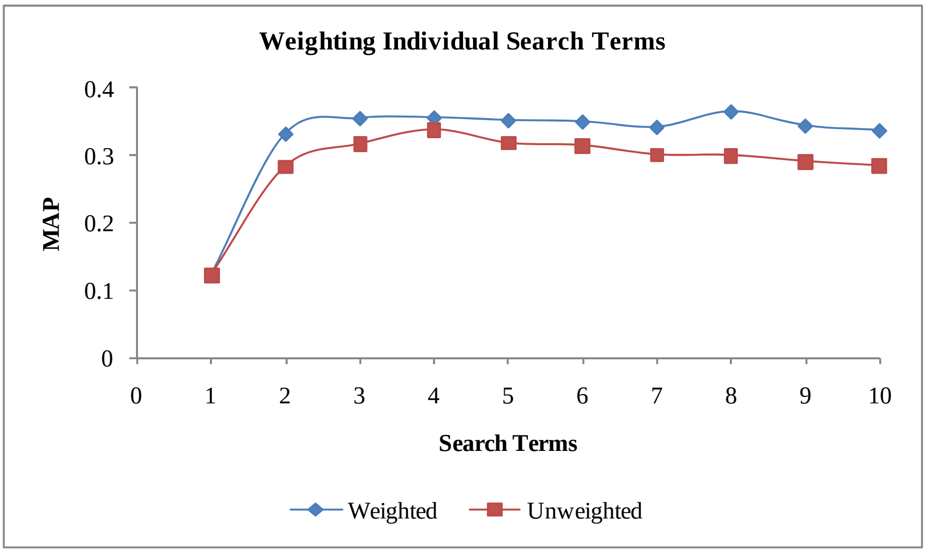

Fig. 5. Effect of weighting individual search terms in the query

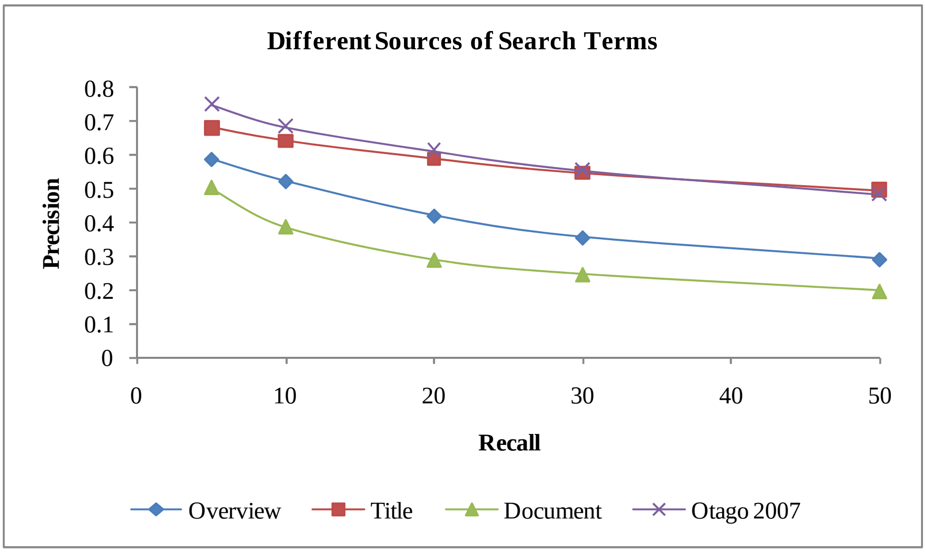

Fig. 6. Different sources of search terms. The title is a more effective source of terms than the overview which is better than the whole document. For early precision the best source was the approach used by Otago at INEX 2007

4. Conclusions

Experiments were conducted to gain insights into effective method of searching for the Relevance-in-Context task. In passage retrieval the Otago 2007 algorithm was compared to the Kullback-Leibler model, and virtually no difference was seen in the performance in the 2008 topics. This suggests the simpler Otago algorithm may be an effective alternative algorithm, especially when efficiency is an issue. In element retrieval the Beigbeder algorithm was compared to an IDF weighted variant and substantial improvements were seen on the 2008 topics – suggesting there is further room for improvement on Beigbeder's work.

In the Link-the-Wiki task the Itakura & Clarke algorithm was used for outgoing links. It was extended by removing punctuation from the anchors, and adding case sensitivity weighting. For incoming links an analysis of the Otago 2007 algorithm suggested that the method of just using the orphan title was effective. At time of writing the results for the Link-the-Wiki track had not been released.

Acknowledgements

Funded in part by a University of Otago Research Grant.

References

1. Jenkinson, D., Trotman, A.: Wikipedia Ad Hoc Passage Retrieval and Wikipedia Document Linking. Focused Access to XML Documents: Proceedings of INEX 2007 (2007) 426-439
2. Huang, W., Trotman, A., O'Keefe, R.A.: Element Retrieval Using a Passage Retrieval Approach. Australian Journal of Intelligent Information Processing Systems **9** (2006) 80-83
3. Itakura, K., Clarke, C.: From Passages into Elements in XML Retrieval. In: Trotman, A., Geva, S., Kamps, J. (eds.): SIGIR 2007 Workshop on Focused Retrieval (2007) 17-22
4. Beigbeder, M.: ENSM-SE at INEX 2007: Scoring with proximity. Preproceedings of INEX 2007 (2007) 53-55
5. Fuhr, N., Kamps, J., Lalmas, M., Malik, S., Trotman, A.: Overview of the INEX 2007 Ad Hoc Track. Focused Access to XML Documents: Proceedings of INEX 2007 (2007) 1-23
6. Itakura, K.Y., Clarke, C.L.: University of Waterloo at INEX2007: Adhoc and Link-the-Wiki Tracks. Focused Access to XML Documents: Proceedings of INEX 2007 (2007) 417-425
7. Geva, S.: GPX: Ad-Hoc Queries and Automated Link Discovery in the Wikipedia. Focused Access to XML Documents: Proceedings of INEX 2007 (2007) 404-416
8. Kamps, J., Koolen, M., Lalmas, M.: Where to Start Reading a Textual XML Document? : 30th SIGIR (2007)
9. Holland, J.H.: Adaptation In Natural and Artificial Systems. Univ. Michigan Press (1975)
10. Trotman, A.: Choosing Document Structure Weights. IP&M **41** (2005) 243-264
11. Robertson, S., Zaragoza, H., Taylor, M.: Simple BM25 extension to multiple weighted fields. 13th CIKM. (2004) 42-49
12. Fachry, K.N., Kamps, J., Koolen, M., Zhang, J.: Using and Detecting Links in Wikipedia. Focused Access to XML Documents: Proceedings of INEX 2007 (2007) 388-403