Using Pertainyms to Improve Passage Retrieval for Questions Requesting Information About a Location

Mark A. Greenwood
Department of Computer Science
University of Sheffield
Regent Court, Portobello Road
Sheffield S1 4DP UK
m.greenwood@dcs.shef.ac.uk

ABSTRACT

This paper explores a method of query formulation for the expansion of natural language questions requesting information about a location, such as *"What is the literacy rate in Cuba?"*. The questions are expanded to form standard information retrieval queries using location pertainym relationships mined from WordNet. Results over the relevant questions from the Text REtrieval Conference (TREC) question answering test sets suggest that selective application of this method produces significantly better performance than using the unaltered questions as queries to an information retrieval engine.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation (efficiency and effectiveness)*

General Terms

Experimentation, Performance, Measurement

Keywords

Question answering, Information retrieval

1. INTRODUCTION

Most open domain question answering (QA) systems contain an initial information retrieval (IR) component which acts as a filter between the full document collection, from which answers are to be drawn, and the detailed processing components required for answer extraction. Such a filter is required as the answer extraction components process text at a level of detail which makes applying them to large text collections impractical.

Clearly, the performance of such an IR component places an upper bound on the end-to-end performance of QA systems – if an IR engine does not retrieve any relevant documents no amount of further processing will enable the question to be correctly answered.

A number of studies [4, 10] have shown that standard IR engines (MG[16] and Okapi[11] respectively) often fail to find answer bearing documents (or passages) when presented with natural language questions. Both studies reported similar results; with between 50% and 60% of questions having at least one relevant document within the top 10 results which increases to almost 80% within the top 100 results. Even when the top 1000 results are considered 8% of the questions have no relevant documents retrieved for them [4] (excluded from this are those questions which are known not to have any relevant documents within the collection).

Unfortunately, it appears that while increasing the number of documents retrieved by an IR engine increases the number of questions for which relevant documents are found, it does not automatically result in an increase in the end-to-end performance of QA systems. This is assumed to be due to the fact that as the volume of text increases the amount of noise (the number of incorrect entities of the correct answer type) also increases. As the noise increases there is a greater chance of the answer extraction components being distracted away from the correct answer. In one recent study [1] the end-to-end performance of a QA system was shown to peak when using just the top 20 passages. These passages contained answers to only 54% of the question set, bounding the performance of the QA system to at most being able to answer 54% of the question set.

These results suggest that what is required is a method which increases the number of questions for which at least one relevant document is retrieved without increasing the volume of text retrieved per question.

Having examined a number of questions and documents known to be relevant, it was noted that whilst a large number of questions include locations, i.e. Q1517 *"What is the state bird of Alaska?"* the answers frequently occur with the adjective form of the location, i.e. *"... willow ptarmigans (the quail-like Alaskan state bird)"* and without the noun form appearing within the relevant passages. Most other words in the question, however, appear unaltered in the relevant passages.

This obvious mismatch between the question and answer texts is likely to mean that relevant documents are either not retrieved or are lowly ranked by most IR engines. Even those IR systems which employ stemming in an attempt to retrieve documents containing morphological variants of question words are unlikely to fare any better as most adjective forms of a location do not produce the same token when stemmed as the stemmed noun – a notable exception

being *Philippines* and *Philippine* which are both stemmed to `philippin` using the well known Porter stemmer[8].

WordNet[6] contains pertainym relations which link adjectives with their associated nouns (i.e. *Alaskan* to *Alaska*) and mining this information allows us to determine the inverse mapping from nouns to adjectives. Together these mappings allows us to experiment with expanding both location nouns and adjectives (i.e. *Alaska* can be expanded to *Alaskan* and vice versa) to form IR queries from natural language questions.

2. RELATED WORK

Numerous studies have reported on methods for query formulation/expansion with respect to natural language question answering. These approaches usually fall into one of the following categories:

1. Queries can be formed from the question words and related words or concepts, i.e. synonyms or morphological variants.

2. Queries can be formed from the question words and terms likely to co-occur with instances of the expected answer type.

An example of the second approach to query formulation is presented in [7]. Questions which are expected to have measurements as answers are expanded to form IR queries by including the associated measurement units, for example:

Q1420[1] *"How high is Mount Kinabalu?"* becomes

```
mount kinabalu alt(meter,inch,foot,centimet)
```

Results of these experiments show significant improvements in the percentage of questions for which at least one relevant document is retrieved – relative improvements of up to 45% over plain IR queries such as `mount kinabalu`.

The main problem associated with expanding questions based upon their expected answer type is that not all possible answer types have clearly defined terms which can be used for expansion. Unlike measurements, answer types such as dates, locations and names have no clearly associated terms which can be used to expand the question to produce an IR query. In these situations, systems have only the question terms (and related knowledge) with which to form a query.

A number of systems [3, 4] expand question terms in order to provide morphological variants and/or synonyms as part of the resulting IR query. Often this leads to complicated queries which are relaxed or constrained through a number of IR iterations before finally arriving at a set of documents that can be processed by the remainder of the system. These systems usually rely on WordNet to expand the question terms, for example:

Q209 *"Who invented the paper clip?"* becomes

```
paper & clip & (invented | inventor | invent)
```

Two issues arise from this approach; word sense disambiguation and common words.

If a question word has more than one sense in WordNet then it has to be disambiguated before synonym expansion can be applied. In the context of ad hoc retrieval it has been shown that the quality of automatic word sense disambiguation (WSD) has a strong impact on retrieval performance [12] (i.e. unless the WSD is accurate any expansion or retrieval will be carried out using irrelevant terms) and it would be logical to assume that the same will apply in the context of question answering.

The second problem arises from the fact that the synonyms of some words (through slang or laziness) are often common words. For example using WordNet to expand the compound term *high school* results in a query including the single word *high*, a relatively common word, which makes the original term *high school* less significant in the resultant query[4].

Both of these problems arise from the fact that query expansion is usually carried out in a brute-force fashion, i.e. expand everything to the full extent possible. A possible solution, therefore, may be to develop a number of selective approaches to query expansion which would improve IR performance while avoiding these pitfalls.

The remainder of this paper presents one such selective approach to query formulation, namely the expansion of location names (and their associated adjectives) using pertainym relationships mined from WordNet[2].

3. EXPERIMENTAL DESIGN

3.1 The Question Set

Two question sets were compiled from the questions prepared for the TREC 11 and 12 question answering evaluations (see [14] and [15] for an overview of these evaluations) and only questions known to have at least one answer within the AQUAINT collection were considered.

The first test set consists of 57 questions containing the name of a country or state and for which a relationship to an adjective can be mined from WordNet. Questions in which the country or state appears as part of a compound noun, for instance Q1753 *"When was the Vietnam Veterans Memorial in Washington, D.C. built?"* were not used. Examples of the questions in this set are:

Q1447 *"What is the capital of Syria?"*
Q1507 *"What is the national anthem in England?"*
Q1585 *"What is the chief religion for Peru?"*

The second question set consists of 31 questions which contain a country or state adjective and for which a pertainym relationship exists in WordNet, examples are:

Q1710 *"What are the colors of the Italian flag?"*
Q1724 *"Who was one of the Egyptian gods?"*
Q2313 *"What does an English stone equal?"*

Together these two question sets allow us to explore the effects on retrieval performance of both expanding locations nouns with their adjective forms and vice versa.

[1]The question numbers relate to the question sets used for the question answering track held as part of the annual Text REtrieval Conference, see `http://trec.nist.gov`

[2]Thanks to Ken Litkowski and Eric Kafe for providing a list of the pertainym relationships found in WordNet

3.2 What Makes a Passage Relevant?

Previous studies have shown that question answering systems often perform better when presented with short passages rather than full documents [9]. Evaluating passage retrieval is, however, more complex than evaluating document retrieval as not only must we determine if the document is relevant but if the system has selected a relevant passage from the document.

Fortunately along with the questions used at TREC, NIST supply a list of relevant documents and a set of regular expressions which match against the known answers for each question[3]. Together these two resources can be used to determine if a passage is relevant by ensuring that not only does it come from a relevant document but that it also matches one of the associated patterns. All the experiments detailed in this paper determine if a given passage is relevant or not using this simple technique.

3.3 Evaluation Metrics

Experiments detailed in this paper will be evaluated using a metric known as coverage (for more details see [10]).

Let Q be the question set, D the document (or passage) collection, $A_{D,q}$ the subset of D containing correct answers to $q \in Q$, and $R^S_{D,q,n}$ be the top n ranked documents in D retrieved by a search engine S given question q.

The *coverage* of a search engine S for a question set Q and document collection D at rank n is defined as:

$$coverage^S(Q, D, n) \equiv \frac{|\{q \in Q | R^S_{D,q,n} \cap A_{D,q} \neq \emptyset\}|}{|Q|}$$

Coverage gives the proportion of the question set for which a correct answer can be found within the top n documents retrieved by S for each question. Note that coverage, unlike the traditional IR measures of precision and recall, does not require that the exact number of relevant documents within the collection be known, only which of the retrieved documents are relevant.

When comparing different approaches to a problem it is important not just to show that one system gives better results than an another but whether or not the differences between the approaches are significant and not due to random chance. Experimental results in this paper are compared using the paired t test [5]. Improvements in the results which are significantly different with 99% confidence are signaled by ▲ while △ signifies only 95% confidence. Similar meanings are attached to ▼ and ▽.

4. QUERY FORMULATION

The experiments detailed in this paper retrieve passages from the AQUAINT collection using the Lucene[4] search engine. Splitting the documents into passages is carried out during index creation and occurs at the paragraph boundaries marked in the source texts. The index was generated using both stopword removal and stemming (Lucene uses the Porter stemmer [8]).

[3] `http://trec.nist.gov/data/qa.html`
[4] `http://jakarta.apache.org/lucene/`

The format of the index implies that queries also have to be subjected to stemming and stopword removal, producing queries such as the following (unless otherwise specified query terms are combined with the `or` operator):

Q1447 *"What is the capital of Syria?"* becomes
`capit syria`
Q1585 *"What is the chief religion for Peru?"* becomes
`chief religion peru`

Two different approaches to query expansion are considered in these experiments. Firstly the queries were expanded by simply including both the noun and adjective form of a location as a nested `or` query within the standard IR query, for example:

Q1447 *"What is the capital of Syria?"* becomes
`capit (syria syrian)`

One possible problem with this approach is that a document which contains both *Syria* and *Syrian* will score higher than those which contain only one of the terms. This is an issue as the original premise was that answer bearing documents, to questions containing the noun form of a location, frequently only contain the adjective form.

To overcome this a new `alt` (alternate) operator was added to Lucene. This operator treats all the terms as alternative versions of the same term (the score is taken to be that of the first term in the `alt` expression). This gives rise to queries such as:

Q1447 *"What is the capital of Syria?"* becomes
`capit alt(syria, syrian)`

This approach treats documents which contain a single instance of *Syria* or *Syrian* in the same way while still ranking documents which contain multiple instances of either form higher than those containing a single instance.

5. RESULTS

Two separate evaluations were carried out to determine the performance benefits of expanding queries using location pertainyms mined from WordNet. The result of expanding queries to include adjective forms of locations contained in the original questions can be seen in Table 1 and Figure 1.

It should be clear from these results that the coverage of the retrieved documents increases when the question is expanded to include the adjective forms of the locations using the `alt` operator. The difference is, however, only significant when we consider 30 or more documents, although this could partly be due to the relatively small size of the question set (only 57 questions). It is also obvious from these results that using the standard `or` operator to expand the queries has a severe detrimental effect on the results. As has already been discussed this is mostly likely to be due to the fact that answer bearing passages tend to contain only one form of the location and using a ranking system that prefers documents which contain both forms pushes answer bearing passages much further down the ranking.

The results of the second evaluation to investigate whether or not expanding adjective locations in questions to include the actual location has an appreciably benefit on the coverage of the retrieved documents can be seen in Table 2 and Figure 2.

Table 1: Coverage results for location noun expansion.

	% Coverage at Rank							
Query Type	**1**	**5**	**10**	**20**	**30**	**50**	**100**	**200**
Question	15.8	31.6	42.1	52.6	52.6	59.6	66.7	75.4
`or` Expansion	7.0 ▽	10.5 ▼	12.3 ▼	12.3 ▼	15.8 ▼	17.6 ▼	21.1 ▼	21.1 ▼
`alt` Expansion	15.8	36.8	45.6	57.9	63.2 ▲	68.4 △	73.7	82.5 △

Table 2: Coverage results for location adjective expansion.

	% Coverage at Rank							
Query Type	**1**	**5**	**10**	**20**	**30**	**50**	**100**	**200**
Question	22.6	38.7	54.8	64.5	67.7	80.6	83.9	87.1
`or` Expansion	9.7 ▽	22.6 ▽	25.8 ▼	29.0 ▼	29.0 ▼	32.3 ▼	38.7 ▼	38.7 ▼
`alt` Expansion	19.4	35.5	51.6	61.3	67.7	77.4	80.6	83.9

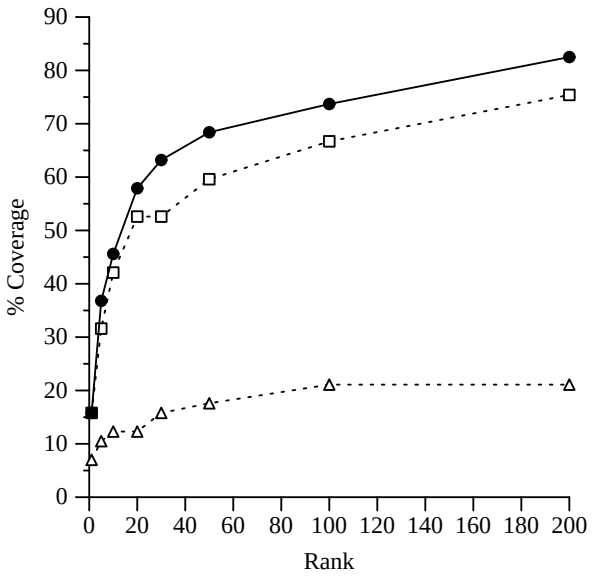

Figure 1: Comparison of standard queries, □, with `alt`, •, and `or`, △, expansion of location nouns.

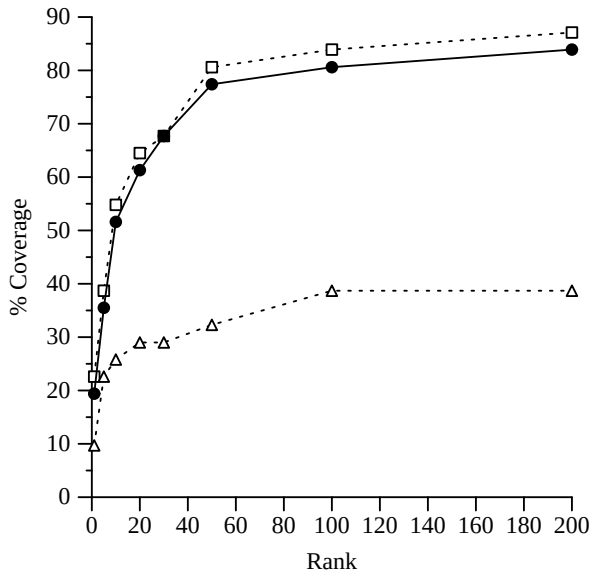

Figure 2: Comparison of standard queries, □, with `alt`, •, and `or`, △, expansion of location adjectives.

This experiment shows that over the current question set the coverage of the retrieved documents is actually reduced when the location is included in the query, although the drop in performance is not significant at any rank examined. A larger test set is required to see if the observed drop in performance is true in general or simply an artifact of the current question set. These results also confirm the results from the first experiment that using the `or` operator to expand the queries has a severe detrimental affect on the performance.

Due to the apparent drop in performance observed in the second experiment when including the location in the IR query a third experiment was undertaken. This third experiment used the first question set (which contains location names) and replaced all the locations with their adjective form rather than expanded to include both forms. The motivation behind this experiment is that while including the adjective form in the first experiment produced an increase in coverage adding the noun form in the second experiment reduced the coverage suggesting that the adjective form may be solely responsible for good IR performance. Queries generated in this experiment include:

Q1634 *"What is the area of Venezuela becomes?"*
`area venezuelan`
Q1647 *"What continent is Scotland in?"*
`contin scottish`

The results of this experiment can be seen in Table 3 and Figure 3. They suggest that the coverage obtained for questions containing locations is dependent upon both the noun and adjective forms of the location and not just the adjective form as seemed to be suggested by the previous experiments.

5.1 Effects on a QA System

Showing that a particular approach to query formulation or expansion increases the coverage of the retrieved documents does not automatically imply that a QA system using these documents will show an increase in performance – higher coverage at the IR stage simply implies a higher upper bound on the performance of answer extraction components. To see if the increase in coverage, obtained through the query formulation approach detailed in this paper, has

Table 3: Coverage results for replacing nouns with adjectives.

Query Type	% Coverage at Rank							
	1	**5**	**10**	**20**	**30**	**50**	**100**	**200**
Question	15.8	31.6	42.1	52.6	52.6	59.6	66.7	75.4
Expanded	12.3	29.8	33.3	47.4	50.9	56.1	59.6	61.4 ▽

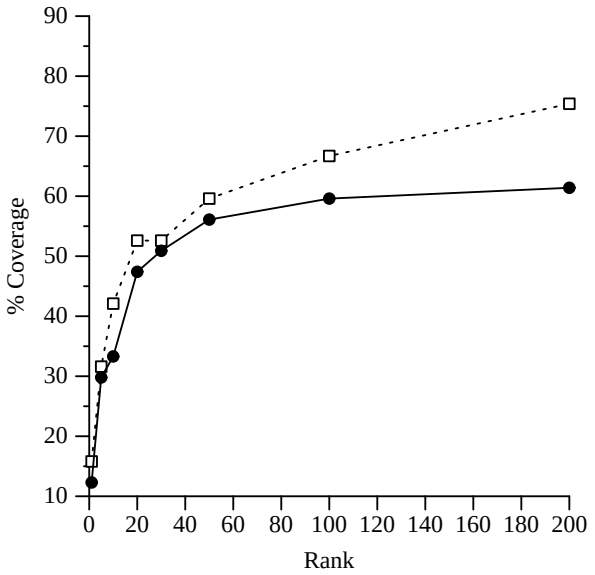

Figure 3: Comparison of standard queries, □, and those in which nouns were replaced by adjectives, •.

a beneficial effect on answer extraction components we provided the retrieved documents as input to an open-domain factoid QA system [2].

The QA system was given as input the top 30 documents (the most significant results were observed when retrieving 30 documents, see Table 1) and was evaluated using MRR (mean reciprocal rank, see [13]) over the top 5 answers returned for each question. The MRR of the QA system when given the documents retrieved using the question alone was 0.1947. Expanding the location nouns in these questions using the `alt` operator resulted in an MRR score of 0.1988. While there is an increase in performance of the answer extraction component it is not large enough to be statistically significant although this could be due to the small size of questions used. Further evaluations over a larger test set are therefore required.

One possible explanation for the small increase in performance may be that while expanding the questions gives better coverage the answer bearing documents can now contain a word (the adjective form of the location noun) which is not part of the question. If the answer extraction components are not adapted to make use of this knowledge then they may well discard answers simply because they appear in a sentence which has little overlap with the original question.

6. CONCLUSIONS

The results of the experiments detailed in this paper show that the original observations motivating the approach were correct: answers to questions which include a location often occur in close proximity to the adjective form of the location, hence including the adjective form in the IR query increases the coverage of the retrieved documents. We also showed that the way in which IR queries are constructed are an important factor in the improvement of the retrieved documents. In the experiments detailed in this paper it is clear that using the `alt` operator gives significantly better results than using the `or` operator. This would not necessarily be the case for all query expansions but can be justified in this context as the original motivation was that often the adjective form of a location appears *instead of* the noun form whereas the `or` operator benefits documents containing both forms.

The results also suggest that the reverse of the original premise does not hold – including the location name in a query when an adjective form appears in the question actually decreases the coverage of the retrieved documents. Further experiments need to be carried out to see if this result is an artifact of the current test set or a more widespread issue.

These experiments also showed that while expanding the questions can give rise to an increase in coverage, of the retrieved documents, due consideration must also be given to improving the answer extraction components to benefit from this expansion.

7. FUTURE WORK

Initially future work should concentrate on building a larger test set to confirm the results obtained in these small scale experiments. This is of special importance for the second experiment (expanding adjectives with their associated nouns) as we failed to find a statistical difference between the results although the coverage of the retrieved documents was reduced at every rank examined.

Another area of future work, should be to investigate if expanding all the nouns in a question with their adjective forms contributes a similar performance increase to expanding just location names. This would, for example, including expanding terms such as *abdomen* to *abdominal* and *volcano* to *volcanic*.

WordNet also contains pertainym relations between adverbs and their stem adjectives. These relationships could also be mined and used in a similar fashion to expand IR queries. For example `abnormally` could be expanded to include *abnormal*.

It is, however, important to remember that simply because one method of query expansion produces an increase in the coverage for a given question set there is no guarantee that a combination of these approaches will produce a retrieval sys-

tem capable of finding relevant documents for all questions asked of it.

From a linguistic point of view one interesting question arises out of the apparent asymmetry in the results of these experiments. Is there some underlying reason why when a noun is used in a question the answer often appears along with the adjective form while the opposite does not appear to follow. Clearly this asymmetry, if true in general, will be of interest to other areas of linguistic study and should be investigated further.

Any future work should also include adapting an answer extraction component to take into account the fact that answer bearing documents may now include highly relevant terms not present in the original question. For example Q1447 *"What is the capital of Syria?"* could now be answered by the text *"... in the Syrian capital, Damascus, ..."* and the answer extraction component needs to be aware that *Syria* and *Syrian* are to be considered equivalent when extracting and scoring *Damascus* as a possible answer.

8. ACKNOWLEDGEMENTS

My thanks to Professor Robert Gaizauskas and Dr Bryony Edwards for their comments and feedback during the preparation of this paper. Any remaining errors or omissions are, of course, my own.

9. REFERENCES

[1] R. Gaizauskas, M. A. Greenwood, M. Hepple, I. Roberts, H. Saggion, and M. Sargaison. The University of Sheffield's TREC 2003 Q&A Experiments. In *Proceedings of the 12th Text REtrieval Conference*, 2003.

[2] M. A. Greenwood and H. Saggion. A Pattern Based Approach to Answering Factoid, List and Definition Questions. In *Proceedings of the 7th RIAO Conference (RIAO 2004)*, Avignon, France, Apr. 27 2004.

[3] S. Harabagiu, D. Moldovan, M. Paşca, R. Mihalcea, M. Surdeanu, R. Bunescu, R. Gîrju, V. Rus, and P. Morărescu. FALCON: Boosting Knowledge for Answer Engines. In *Proceedings of the 9th Text REtrieval Conference*, 2000.

[4] E. Hovy, L. Gerber, U. Hermjakob, M. Junk, and C.-Y. Lin. Question Answering in Webclopedia. In *Proceedings of the 9th Text REtrieval Conference*, 2000.

[5] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 2000.

[6] G. A. Miller. WordNet: A Lexical Database. *Communications of the ACM*, 38(11):39–41, Nov. 1995.

[7] C. Monz. *From Document Retrieval to Question Answering*. PhD thesis, Institute for Logic, Language and Computation, University of Amsterdam, 2003. Available, April 2004, from `http://www.illc.uva.nl/Publications/Dissertations/DS-2003-04.text.pdf`.

[8] M. Porter. An Algorithm for Suffix Stripping. *Program*, 14(3):130–137, 1980.

[9] I. Roberts. Information Retrieval for Question Answering. MSc Dissertation, Department of Computer Science, The University of Sheffield, UK. Available, Februray 2003, from `http://www.dcs.shef.ac.uk/teaching/eproj/msc2002/abs/m1ir.htm`, 2002.

[10] I. Roberts and R. Gaizauskas. Evaluating Passage Retrieval Approaches for Question Answering. In *Proceedings of 26th European Conference on Information Retrieval*, 2004.

[11] S. E. Robertson and S. Walker. Okapi/Keenbow at TREC-8. In *Proceedings of the 8th Text REtrieval Conference*, 1999.

[12] M. Sanderson. Retrieving with Good Sense. *Information Retrieval*, 2(1):49–69, 2000.

[13] E. M. Voorhees. Overview of the TREC 2001 Question Answering Track. In *Proceedings of the 10th Text REtrieval Conference*, 2001.

[14] E. M. Voorhees. Overview of the TREC 2002 Question Answering Track. In *Proceedings of the 11th Text REtrieval Conference*, 2002.

[15] E. M. Voorhees. Overview of the TREC 2003 Question Answering Track. In *Proceedings of the 12th Text REtrieval Conference*, 2002.

[16] I. H. Witten, A. Moffat, and T. C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann, second edition, 1999.