Webology, Volume 3, Number 4, December, 2006

More Effective Web Search Using Bigrams and Trigrams

David Johnson
PhD Candidate, School of Computing, Private Bag100, University of Tasmania, Hobart, Tasmania 7000, Australia. E-mail: dgjohnso (at) utas.edu.au

Vishv Malhotra
Senior Lecturer, School of Computing, Private Bag100, University of Tasmania, Hobart, Tasmania 7000, Australia. E-mail: vishv.malhotra (at) utas.edu.au

Peter Vamplew
Senior Lecturer, School of Information Technology and Mathematical Sciences, University of Ballarat, P.O. Box 663, Ballarat, Victoria 3353, Australia.
E-mail: p.vamplew (at) ballarat.edu.au

Received October 10, 2006; Accepted December 13, 2006

Abstract

This paper investigates the effectiveness of quoted bigrams and trigrams as query terms to target web search. Prior research in this area has largely focused on static corpora each containing only a few million documents, and has reported mixed (usually negative) results. We investigate the bigram/trigram extraction problem and present an extraction algorithm that shows promising results when applied to real-time web search. We also present a prototype augmented search software package that can leverage the results provided by a web search engine to assist the web searcher identify important phrases and related documents quickly. This software has received favourable feedback in a recent user survey.

Keywords

Web searching; Information need; Relevance feedback; Part-of-speech tagging

---

Introduction

Quickly and easily finding required information on the Web remains a major problem, particularly in domains where the searcher has little prior knowledge.

This is despite significant improvements in search engine technology in recent times.

One reason for this problem is the exponential growth of available web resources since the early 1990s. The Internet Systems Consortium (2006) reports that as at July 2006 there were some 439 million Internet hosts, and Sullivan (2004) reports that Google claims to index over 8.1 billion web pages, with MSN close behind at 5.0 billion and Yahoo at 4.2 billion.

Another reason is that commercial web sites put significant effort into attracting traffic from web search engines by developing pages that appear relevant to search queries, but are in fact nothing more than collections of links, keywords and largely "attention-grabbing" text often created by automated "content generation" programs.

A third reason is the inherent ambiguity of human language. Most words have more than one possible meaning (polysemy) and there are also usually many words that can express the same concept (synonymy). It is thus a difficult problem to determine if a particular search keyword that appears on a web page is being used to express the concept that the searcher is interested in, or on the contrary, if a word that is completely different from the keyword is indeed relevant. The problem is further confounded by the variety of spellings used around the world, and the numerous misspellings that abound on the Web.

Web searchers often make use of short quoted phrases (using an "exact phrase" search option or placing the phrase in quotation marks) to quickly locate known items. Often three or four words quoted from the title or text of a document is sufficient to locate it uniquely (along with other documents that quote it). This seems to be a tactic used by many web searchers as our analysis of some 5.7 million web search queries from publicly released 2002 AltaVista and AlltheWeb query logs (Jansen, 2006) found that some 11.7% of all queries contained search terms in quotation marks. The majority of these (78.1%) were two and three word phrases (bigrams and trigrams).

Although web searchers appear to recognize the value of phrase searching in effectively targeting information, they usually have little to guide them in selecting suitable phrases, particularly in unfamiliar domains. In this research, we have investigated methods that leverage the results obtained from standard search engines to assist a searcher in selecting search phrases. We have also investigated the use of extracted phrases as an "index" to browse a document collection arising from an initial search query, and evaluated the results obtained from augmented queries using quoted phrases.

The remainder of this paper is organized as follows. In the next section, we discuss why phrases should be effective query building blocks, the problems they pose and

introduce our proposed solutions and prototype software. In the evaluation section, we discuss the results from two different experiments, and then move on to our conclusions and proposals for further work.

Discussion

Multi-word features such as bigrams and trigrams convey more specific meaning than single word features, and therefore should be more effective in targeting relevant web search results. For example, the word "mole" has some 20 meanings listed in the "Mole disambiguation page" (Wikipedia, 2006) - presented with the word "mole" in isolation, we have little other than general usage frequency to guide us about what sense of the word is relevant, but in the context of an appropriate bigram or trigram, the intended sense becomes much clearer, for example "facial mole", "Adrian Mole", "mole sauce" and "undercover enemy mole" all have distinct and clear meanings.

Previous studies into the effectiveness of multi-word features in information retrieval (IR) systems have reported mixed results - sometimes a small improvement in retrieval effectiveness is achieved, and often no improvement or even degradation in performance is found. Some examples are Croft et al., 1991; Lewis & Croft, 1990; and Strzalkowski & Carballo, 1997. Two major problems encountered with multi-word features in traditional IR are the sparse data problem (discussed in the next section), which creates difficulties with modelling and estimation, and the problem of identifying bigrams and trigrams that have high semantic content. In this study, we address these problems by only considering phrases extracted from our initial corpus and which match part-of-speech templates that have been found to identify useful phrases.

The sparse data problem

When a searcher enters any query, the goal of an effective IR system is to create a list of documents that most closely match the query, even if some (or all) of the terms in the query do not occur in the corpus. In the case where quoted phrases are used as query terms, the "sparse data problem" arises, where many of the possible query phrases that a searcher could enter will not exist in the corpus.

The extent of the problem is illustrated by considering a vocabulary of 25,000 words. This could generate almost 625 million distinct bigrams and over 15 trillion trigrams. Clearly most of these will not occur in the corpus.

As discussed in Manning & Schutze (2003a), the estimation methods that are used in modern IR systems basically rely on discounting to smooth the data. The probability of seen events (that is phrases that have been found in the corpus) is reduced a little to allow some of the probability mass to be left over to account for as yet unseen events. While these methods do help with the estimation process,

they appear better suited to handling single word features where the underlying data is significantly less sparse, and it is our contention that many of the disappointing results reported from bigram/trigram models have roots in the estimation problem.

In our current work we side-step the sparse data problem by considering only phrases actually found in the corpus that consists of text from documents returned by an initial web query related to the area of interest. The bigrams and trigrams extracted from this corpus will be representative of various aspects of the topic area - some of which will be related to our information need and some of which will not.

The effectiveness of using these to improve the search query will therefore depend to a large extent on being able to extract meaningful phrases from the text, as discussed next.

Extraction of useful bigrams and trigrams

In English text the most commonly occurring bigrams and trigrams convey little meaning. Manning & Schutze (2003b) analysed roughly 14 million words from *New York Times* newswire articles and the 10 most common bigrams found are: "of the", "in the", "to the", "on the", "for the", "and the", "that the", "at the", "to be" and "in a" - not very useful at all. In fact, of the 20 most common bigrams found "New York" (15^{th} on the list) was the only phrase that could be considered to convey useful meaning. Our analysis of text from about 10,000 English language web pages also returned similar results.

In our analysis, most of the low utility bigrams and very few of the high utility bigrams were found to contain at least one "stop word" (a common word such as "a", "of", "the", etc. which in general conveys little meaning - it is a common IR practice to strip such words from text before indexing). The main exception to this observation is where the word "the" is used in front of a proper noun, for instance "The Times" may be a reference to the U.K. based newspaper, although it could be an inconsequential text fragment, "The times recorded were well outside the record . . ."

Simply ignoring bigrams that contained at least one stop word was found to be a very simple and reasonably effective method of discovering meaningful bigrams without losing too many of them, and in fact was the method used for much of our earlier work which did not consider trigrams. Extraction of meaningful trigrams proved more difficult.

Previous work, for example Justeson & Katz, 1995, reported success in extracting high utility phrases (referred therein as "content phrases") from text using part-of-speech (POS) filtering. The POS templates used to identify potentially useful bigrams and trigrams in that work were "Adjective Noun", "Noun Noun",

"Adjective Adjective Noun", "Adjective Noun Noun", "Noun Adjective Noun", "Noun Noun Noun" and "Noun Pronoun Noun".

In order to find a suitable tagger, we evaluated The Stanford NLP Group Tagger (2006), MontyTagger (2006) and QTag (2006). QTag was found to be the fastest (by at least an order of magnitude compared to the slowest) and most robust when dealing with misspelt words and other "junk" text. There was little difference in accuracy between the three. Speed is quite an important consideration since we want to download the text, analyse it, and present the results to a web searcher as soon as possible after they submit their initial query.

In our early experiments, we ran the text of a few thousand web pages (the top 100 documents from each of 100 web searches) through the POS tagging/filtering process (using the seven POS templates mentioned above) and found that while the quality of the accepted phrases was quite high, but too many wanted phrases were rejected. In some cases good phrases were rejected because words were incorrectly tagged, and in other cases useful phrases just didn't fit any of the templates. After some analysis and fine-tuning we ended up with 40 bigram and 61 trigram templates that would (after rejection of phrases containing stop words) accept almost all potential content phrases while keeping junk to a minimum. These templates are listed in Appendix 1.

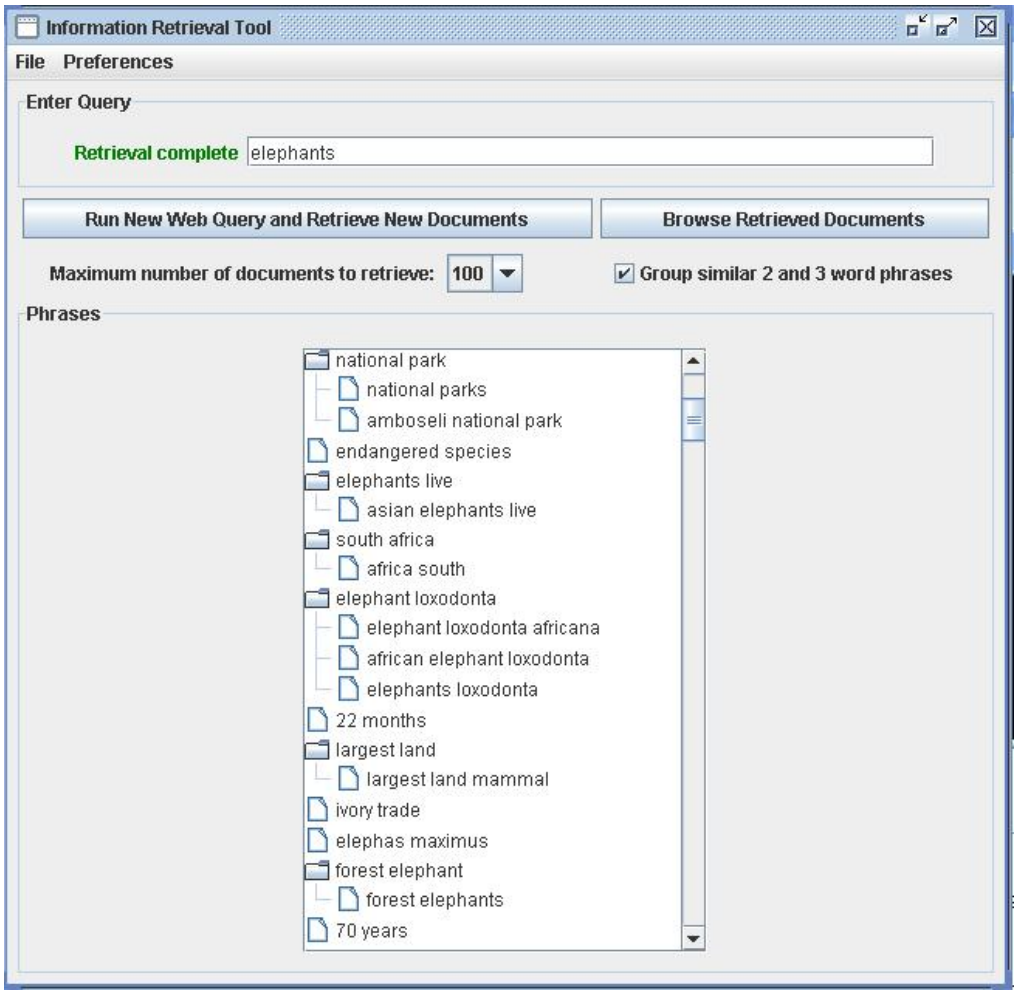
Assisting the Web searcher

We developed an interactive application designed to support the search process and assist the searcher in identifying phrases and documents of interest. This application works as follows:

- An initial query on the topic of interest is entered (usually 2-3 words in our tests, which is the average length of a web search query, as reported by Silverstein et al., 1999; amongst others).
- This is then submitted to a web search engine (Google in our prototype) and the text from the first n (100 for most of our tests) web pages returned by the search is downloaded.
- The text is analysed and the user is presented with a list of the "meaningful" bigrams and trigrams that have been extracted. (These are presented in order of descending frequency, with some additional grouping of similar phrases to increase usability. Also phrases must occur in at least a threshold number of documents, currently set at four, before they are shown.)

By browsing through the list of phrases the searcher can immediately get a feel for the topic areas and salient facts contained in the search results. From this list the searcher can also choose to browse any or all of the documents that contain a particular phrase (or documents that contain the chosen phrase as well as another chosen phrase that commonly co-occurs with it). This is illustrated in Figure 1.

Figure 1: Prototype augmented search application showing part of the phrase list obtained from the initial query "elephants"

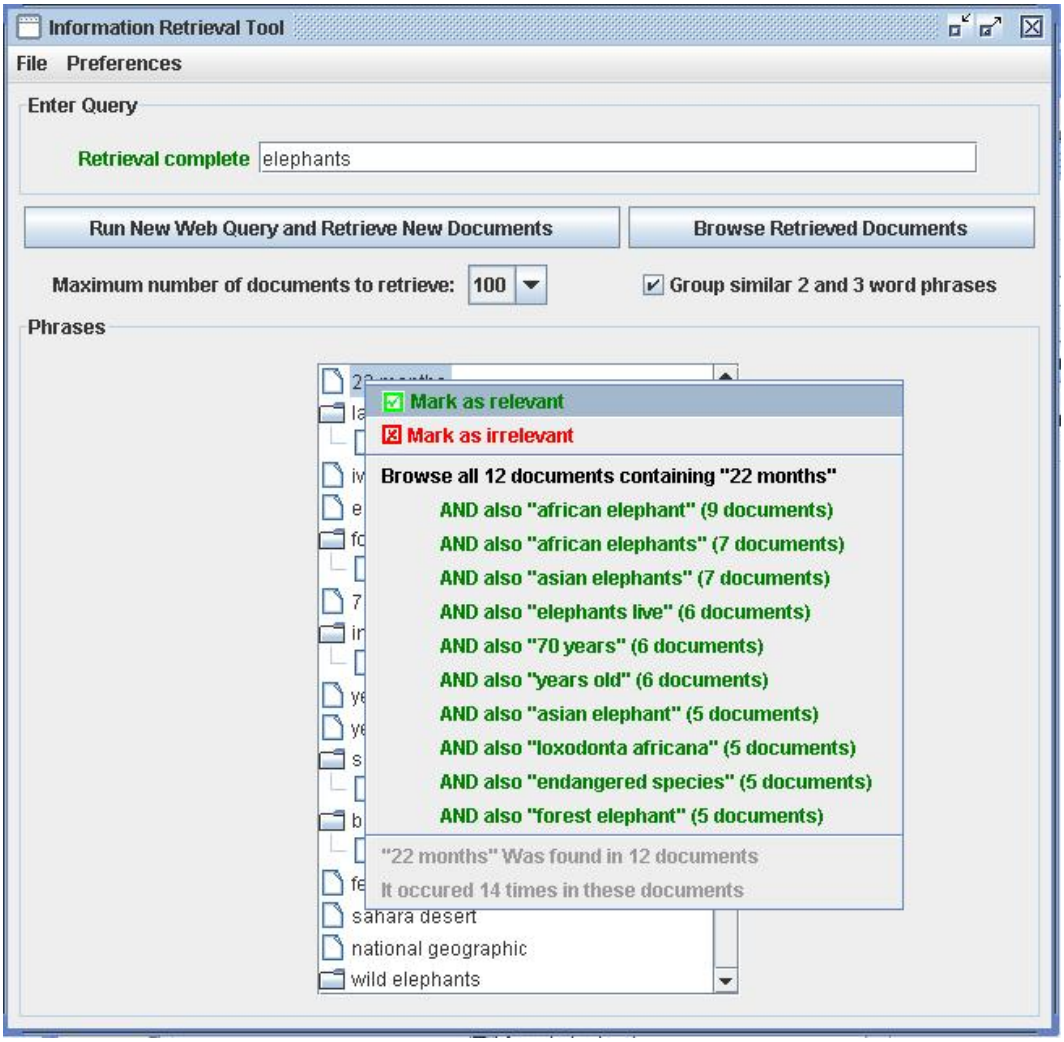


If the searcher feels that the initial search results are inadequate (either by looking through the phrase list or by browsing documents), appropriate phrases can be marked as relevant or irrelevant and a new, more targeted query can be submitted to the search engine. The new query will have the form:

initial query ("relevant phrase 1" OR "relevant phrase 2" . . .) -"irrelevant phrase 1". . .

As a concrete example, suppose a searcher is seeking information about elephants, and starts the search process with the initial query "elephants" as was shown in Figure 1. Amongst the first few phrases extracted from returned documents are "African elephant", "Asian elephant", "loxodonta africana", "national park", "endangered species", "South Africa", "largest land mammal", "forest elephant", "baby elephant", "wild elephants", "ivory trade", "elephant conservation" and "22 months". Browsing this list the searcher immediately gains some insights into the topic areas covered in the document collection that the search has created, and is guided into deciding which phrases and documents are most relevant.

Figure 2: Options available after clicking on a phrase in the phrase list



A searcher interested in the life-cycle of the elephant species may decide to investigate some of the 12 documents that contain the phrase "22 months", since this may remind them that this is likely to be the gestation period of the elephant.

Browsing the document containing the most (3) occurrences of "22 months", they would find ELEPHANTS IN CAPTIVITY, a page containing excellent information on the life-cycle and habits of both African and Asian elephants. This is a page that they would most likely not have found unaided, since it was ranked in 73rd position on the initial search. (The results in this example are based on a search using Google as at 20th September 2006 with the text of the first 100 documents being downloaded).

Evaluation

In this section, we present the results of our evaluation of phrases as search query units, as well as results obtained from a small user survey.

The traditional IR effectiveness measures of precision (number of relevant documents retrieved/number of documents retrieved for various sized retrieval sets) and recall (number of relevant documents retrieved/number of relevant documents available), and the various metrics that combine these two, are not ideal measures of success in real world web searching tasks.

The precision metric can be useful for small retrieval set sizes, for instance looking at the number of relevant documents in the first 10 or 20 search engine results because it is important for the search to return as many relevant results as possible early in the rankings. (As discussed by Jansen et al., 2000, amongst others, over 50% of web searchers do not venture past the 1st page of results - usually at 10 results per page.)

Judging the relevance of documents is, in itself, problematic. Even if we have a comprehensive description of what the relevant information is, how do we treat a document that is mainly discussing a different topic but does provide some useful information on our topic? Also what if we have two documents that cover exactly the same points? We could argue that the second is irrelevant once we have found the first - it may support the veracity of the first document, but it does not add any new information. An even more contentious example is where the information in one document is a subset of the information in another. We could argue that if we saw the smaller document first then they are both relevant as the larger adds new information, but if we saw them in the reverse order the smaller would be irrelevant.

In this study, we have adopted the pragmatic, though imperfect, definition of relevance used in the NIST Text REtrieval Conferences (TREC) - " *TREC uses the following working definition of relevance: If you were writing a report on the subject of the topic and would use the information contained in the document in the report, then the document is relevant. Only binary judgments ("relevant" or "not relevant") are made, and a document is judged relevant if any piece of it is relevant (regardless of how small the piece is in relation to the rest of the document).*" TREC (2006).

Measuring recall is even more problematic since the number of relevant documents available (even on the portion of the Web that is indexed by search engines) is unknown. None of the popular search engines (e.g., Google, Yahoo, AlltheWeb, AltaVista, Excite, MSN, AOL, Ask.com) will return more than 1,000 links for a search, even though they may estimate that there are hundreds of thousands or even millions of pages that match it.

Even the estimates of the number of pages that match a search can be quite misleading. As an example - a Google search for automobile recalls "recalls defects" as at 19th September 2006 reported there were "about 236,000" results from the search, but upon viewing page two of the returned results we see the

message "*In order to show you the most relevant results, we have omitted some entries very similar to the 11 already displayed*". In fact the 11 results came from only seven different domains, and repeating this search with the omitted results included, confirmed that, at least in the first few hundred links all results were indeed very similar, and came from these same seven domains.

Effectiveness of bigrams and trigrams as query phrases

Experiments were conducted to evaluate results obtained from web queries incorporating bigrams and trigrams. Two methods were used to identify suitable phrases to add to the initial query. In method A, the first 10 documents returned from an initial search were examined and each was marked as either relevant or irrelevant to the topic area. The text of these documents was then processed using the standard relevance feedback (RF) equation [as first reported in Rocchio (1971)] to generate two new enhanced queries - the first of which could contain additional single words as well as phrases, and the other containing additional phrases only. In method B, the text of the first 100 documents from the initial search was automatically downloaded and analysed to extract content phrases as previously. The searcher was then presented the list of these phrases and asked to choose suitable phrases to add to the augmented query. For both methods A and B the augmented queries were then resubmitted and the number of relevant documents in the first 10 results used as a measure of performance. These experiments were conducted by a member of our research team as a "proof of concept" study.

The 12 topics were chosen at random from TREC (2004) Robust track topics 304 to 448 and the initial web search was performed using Google, with the title of the topic being the initial search query (for example "Greek Philosophy Stoicism" - submitted without quotation marks). For half of the topics method A was followed by method B, and for the remainder the order was reversed to eliminate the possibility that knowledge gained from the first search on a given topic would systematically affect the second.

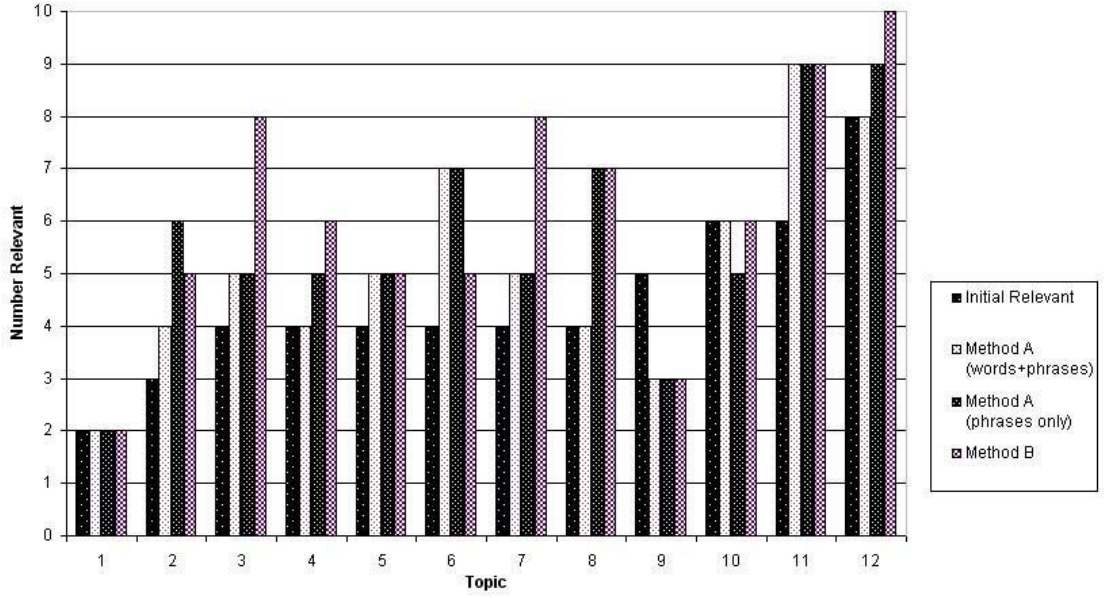
In each case, the time spent by the searcher in either selecting phrases or reviewing documents was measured. The time taken to run the searches and download the documents was not taken into account in these results. The time to download and analyse the text of 100 documents is approximately 30 seconds on a 2.4GHz Pentium 4 PC with 512MB RAM, with a 1.5Mbps Internet connection. This high level performance is achieved by making good use of simultaneous download threads and overlapping the download and analysis tasks as much as possible.

Table 1 summarizes the results obtained, while Figure 3 illustrates the results on individual queries. The minimum number of relevant documents returned in the initial search was 2/10, while the average was 4.58/10.

Table 1: Summary of results - initial queries averaged 4.58 relevant links in the first 10 returned

	Method A (words + phrases)	Method A (phrases only)	Method B
Average searcher evaluation time	702 seconds	702 seconds	94 seconds
Average relevant (1st 10 search links)	5.25	5.66	6.2
Topics with improvement	6	9	9
Topics with degradation	1	2	1

Figure 3: Number of relevant links in first 10 search results for each individual query used in survey



In almost all cases the augmented queries performed better than the initial queries, and the queries augmented with phrases alone performed best, supporting the value of phrases as valuable query components.

The two cases where an improvement on the initial query was not made were topic 1 and topic 9. Topic 1 was "marine vegetation", specifically looking for information regarding commercial harvesting of marine vegetation for food or drug purposes. All pages retrieved from the initial and subsequent searches dealt almost exclusively with conservation and/or marine reserves, with only passing mention of the topic in the two pages judged relevant, thus providing virtually no additional useful query terms. Topic 9 was "abuses of e-mail", specifically relating dissatisfaction of employers to abuses by employees engaging in communications not related to their work. In this case the initial query did quite well with 5/10 relevant documents, but the augmenting terms, such as "civil liberties" and "instant messaging" managed to target the search back to the dominant topic of bulk unsolicited e-mail ("spam").

Of particular interest is that method B resulted in highly effective augmented queries, while also being much faster in terms of the searchers evaluation time than method A. Not all of this time can be claimed as a saving for the searcher, since information on the topic is obviously gained while evaluating pages for method A, however it does provide evidence that the approach taken in our prototype augmented search application can be effective.

User Survey

The aim of the user survey was to compare our augmented search prototype software with a traditional search engine (in this case Google) in regards to the time taken to search the assigned topics, depth of information retrieved, and user satisfaction with the search process. Google was also used as the underlying search engine to obtain the document links for the augmented search prototype.

The survey was conducted in May 2006. A group of 16 volunteer Information Technology students (from grades 11 and 12, aged 16 to 18 years) were each given two search topics to investigate using Google alone and two topics to search with the assistance of our prototype software, during a session of a little under two hours.

The topics were chosen from TREC (2004) Robust Track topics 304 to 448, these being the topics that had been most difficult (producing the lowest precision and recall figures) in previous TREC trials. Participants were briefed to obtain sufficient information to form the basis of a hypothetical short essay assignment of around 2,000 words).

Topics were assigned so that each was searched by one student using the augmented search prototype software and by another student using Google alone to allow a direct comparison of performance with a fixed topic. Relevant links and text were transferred to a separate "Topic Manager" application when found (and they could also be deleted if later search revealed any to be irrelevant). The "Topic Manager" logged when search events occurred, and at the end of each search it presented a short survey to gauge the searcher's satisfaction with various aspects of the search process (on a scale of 0 to 10, higher being better) and collect other comments. The results of the survey are summarized in Table 2.

Table 2: Average ratings of search methods reported in post-search user surveys

Search method	Search effectiveness	Depth of information coverage	New information learnt	Proportion of viewed pages that were useful	Utility of phrase list	Utility of local search page
Google	6.22	6.33	6.89	6.11	NA	NA
Augmented search	6.22	6.44	6.56	6.22	5.89	6.33

Whilst supervising the survey, and subsequently analysing the results, it became apparent that there were too many confounding variables to draw meaningful quantitative conclusions (for instance differences in abilities, attention spans and motivation of the searchers; discrepancies in relevance judgements; some network glitches/performance issues; and some pages being inappropriately blocked by network content censoring software). It did provide some evidence of the merit in our approach, and showed that it at least did not hinder the search process. The survey also provided a good opportunity to see our prototype in use by searchers who were using it for the first time. Almost all of the feedback received was positive; most participants commented that they thought the prototype was useful. Several minor improvements were made to the prototype based on our observations and feedback received during this survey. We are planning to conduct a new survey using a different evaluation methodology in the near future and are hopeful of obtaining more conclusive results.

Conclusions

Quoted bigrams and trigrams that have been extracted from a corpus derived from the text of documents from an initial web query can provide a good basis for augmenting simple web queries to provide more targeted results. The results support our contention, discussed in the "Sparse data problem" section, that many of the disappointing results reported from bigram/trigram IR models have roots in the estimation problem, rather than inherent problems with the use of bigrams and trigrams.

We have found that useful ("content") phrases can usefully be extracted from such a corpus by means of relatively straightforward part-of-speech tagging and template matching.

The extracted phrases have been shown to provide a means to quickly locate and browse relevant documents in the corpus. These phrases have been found more useful than single words in this context due to their higher semantic content.

We have also found that a searcher can work directly from an automatically extracted phrase list to formulate an augmented query that will in most cases perform better than the initial query (and similarly, if not better than a query generated by mechanical relevance feedback algorithms), without needing to read and review the underlying documents. This can result in significant savings of time and cognitive effort (particularly when many irrelevant pages are involved).

Future work

A further user study is planned, taking a more user-centred approach, loosely based on the approach outlined by Spink (2002). A fixed time will be allowed for searching, after which a post-search questionnaire will be given to gauge the

change in the searchers personal knowledge on the topic area as a result of the search. It is planned to have a series of factual questions relating to each topic, and the searcher will be asked to assess for each question if their answer is based on prior knowledge or information gained while searching.

We are also evaluating alternative user interfaces for our prototype software and we plan to evaluate the feasibility of incorporating its functionality into a server-based application that could be accessed by a searcher in a similar manner to existing search engines.

Acknowledgements

We would like to thank Ms. Bobby Court (Principal), Mr. Jeremy Dooley, Mr. Donnie Roland and the students of Guilford Young College Glenorchy for their invaluable assistance and participation in the user survey. Permission to conduct a human survey was approved by the Human Research Ethics Committee (Tasmania) in their notice H8874 dated 12th May 2006.

References

- Croft, W.B., Turtle, H.R., & Lewis, D.D. (1991). The use of phrases and structured queries in information retrieval. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp 32-45).
- Internet Systems Consortium (2006). ISC Internet Domain Survey. Retrieved September 15, 2006, from http://www.isc.org/index.pl?/ops/ds/
- Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real life, real users and real needs: A study and analysis of users' queries on the Web. *Information Processing and Management*, 36(2) (pp 207-227).
- Jansen, J. (2006). AltaVista, AllTheWeb and Excite search engine query logs generously made available by Dr Jim Jansen. Retrieved September 15, 2006, from http://ist.psu.edu/faculty_pages/jjansen/
- Justeson, J.S., & Katz, S.M. (1995). *Technical terminology: Some linguistic properties and an algorithm for identification in text.* Natural Language Engineering, (pp. 9-27). Cambridge University Press.
- Lewis, D.D., & Croft, W.B. (1990). Term Clustering of Syntactic Phrases. *Proceedings of the Thirteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.
- Manning, C.D., & Schütze, H. (2003a). *Foundations of Statistical Natural Language Processing* (sixth printing), (pp. 196-217). The MIT Press.
- Manning, C.D., & Schütze, H. (2003b). *Foundations of Statistical Natural Language Processing* (sixth printing), (pp. 29-34). The MIT Press.
- MontyTagger (2006). The MontyLingua natural language package. Retrieved June 5, 2006, from http://web.media.mit.edu/~hugo/montylingua/index.html

- QTag (2006). QTag probabilistic parts-of-speech tagger. Retrieved June 5, 2006, from http://www.english.bham.ac.uk/staff/omason/software/qtag.html
- Rocchio, J. (1971). Relevance feedback in information retrieval. In Salton, G. (Ed.), *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice-Hall, (pp. 313-323).
- Silverstein, C., Henzinger, M., Marais, H., & Moricz, M. (1999). Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1), 6-12.
- Spink, A. (2002). A user-centered approach to evaluating human interactions with web search engines: an exploratory study. *Information Processing and Management* , 38(3), 401-426.
- The Stanford NLP Group Tagger (2006). The Stanford NLP Group Log-linear Part-Of-Speech Tagger. Retrieved June 5, 2006, from http://nlp.stanford.edu/software/tagger.shtml
- Strzalkowski, T., & Carballo, J.P. (1997). Natural Language Information Retrieval: TREC-4 Report. In Harman, D. (Ed.), *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*. Washington, D.C.
- Sullivan, D. (2004). Search Engine Size Wars V Erupts. *SearchEngineWatch*, Retrieved on September 10, 2006, from http://blog.searchenginewatch.com/blog/041111-084221
- TREC (2004). Robust Track: Robust test set. Retrieved from http://trec.nist.gov/data/t13_robust.html
- TREC (2006). Text Retrieval Conference: Data - English Relevance Judgments. Retrieved August 8, 2006, from http://trec.nist.gov/data/reljudge_eng.html
- Wikipedia (2006). "Mole disambiguation page". Retrieved September 1, 2006, from http://en.wikipedia.org/wiki/Mole

---

Appendix 1 - Part-of-speech tags used to identify potential content phrases

Key to POS tags:

C = conjunction
D = determiner
F = foreign word
I = preposition
J = adjective
M = modal auxiliary (might, will)
N = noun
O = ordinal number
P = pronoun
R = adverb
S = symbol or formula
V = verb
W = possessive pronoun

Bigram Tags
CJ, CN, CR, CV, DJ, DN, DV, FN, IC, IN, IR, IV, JC, JF, JJ, JN, JR, JV, NC, NF, NI, NJ, NN, NR, NS, NV, NW, OC, OJ, ON, PJ, RJ, RN, RV, VD, VI, VJ, VN, VS, WN

Trigram Tags
DJN, DJV, DNC, DNN, DNV, DRN, ICJ, ICN, IJN, INJ, INN, INR, INV, IRN, IRR, IVJ, IVN, IVV, JCN, JFN, JJJ, JJN, JNI, JNN, JNV, JON, JRN, JVI, JVJ, JVN, JVV, MVN, MVV, NFN, NFV, NIN, NIV, NJN, NJV, NNC, NNF, NNJ, NNN, NNV, NOJ, NVC, NVN, OJN, ONN, PJN, RJN, RNN, RRN, RVJ, RVN, VCN, VJN, VNN, VRC, VVJ, VVN

Bibliographic information of this paper for citing:

Johnson, David, Malhotra, Vishv, & Vamplew, Peter (2006). "More Effective Web Search Using Bigrams and Trigrams." *Webology*, **3**(4), Article 35. Available at: http://www.webology.ir/2006/v3n4/a35.html

---
