AUGMENTING SEARCH WITH CORPUS-DERIVED SEMANTIC RELEVANCE

Zachary Mason
Brandeis University
Waltham, MA

Keywords: Corpus linguistics, semantic search, query refinement, semantic modeling.

Abstract: This paper describes a system for doing contextually-steered web search. The system is based on a method for estimating the semantic relevance of a web page to a query.
Consider doing a web search for conferences about web search. The query "search conferences" is not effective, as it produces results relevant for the most part to searching over conferences, rather than conferences on the topic of search. The system described in this paper enables queries of the form *"search conference context:pagerank"*. The *context* field in this example specifies a preference for results semantically relevant to the term "pagerank", although there is no requirement that said results contain the word "pagerank" itself. This a more semantic, less lexical way of refining the query than adding literal conjuncts.
Contextual search, as implemented in this paper, is based on the Google (Google) search engine. For each query, the top one hundred search results are fetched from Google and sorted according to their relevance to the context query. Relevance is computed as a distance function between the vocabulary vectors associated with a web-page and a query. For queries, the vocabulary vector is formed by aggregating the web-pages in the search results for that query. For web-pages, the vocabulary vector is aggregated from that web-page and other web-pages nearby in link-space.

1 SEMANTIC MODELS OF TEXT

The Internet has a tremendous amount of information much of which is encoded in natural language. Human natural language is innately highly polysemous at both the word and phrasal level, so texts are rife with ambiguity. This is a problem for purely lexical search engines. One can refine an ambiguous query by successively adding qualifiers, but this can be time consuming and the variety of ways a given idea can be expressed can make the addition of query conjuncts dangerously restrictive.

For contextual search we need a way to computationally model the semantics of short texts - queries are usually no more than a few words and the amount of text on a web page can be as low as zero. What is needed is an approach that supports quick computations and requires no background knowledge. In the approach described in this paper, the semantic representation need only support a similarity operator (it is not necessary that, for instance, propositional information should be extractable from it.) Further requirements are that representations should be compact, should be noise tolerant, and should permit the comparison of arbitrary texts. Our solution is to use vectors of associated vocabulary to model the semantics of queries and web pages.

For a query, we obtain a vocabulary vector by doing a web-search on that query (on (Google), for instance), taking all the snippets associated with each of the top 100 search results and breaking them into a bag-of-words representation.

A more thorough approach is fetching N result links from the web search, follow them, and amalgamating their text. The disadvantage to this approach is the time required - web-pages may be served slowly, in practice averaging on the order of seconds to load, and in any event this approach is bandwidth intensive. Empirically, we find that the expanded representation obtained from using whole web pages rather than snippets does not improve performance (probably because with snippets performance is already very

Mason Z. (2007).
AUGMENTING SEARCH WITH CORPUS-DERIVED SEMANTIC RELEVANCE.
In *Proceedings of the Third International Conference on Web Information Systems and Technologies - Web Interfaces and Applications*, pages 367-371
DOI: 10.5220/0001259403670371

Table 1: Most frequent symbols from vocab. vector of query “pagerank”.

count	symbol
15	software
14	tutorials
12	technology
11	programming
11	development
9	applets
7	articles
6	project
6	enterprise
6	edition
6	developers
6	comprehensive
6	books
5	virtual
5	training

Table 2: Most significant unigrams for “William Gibson”.

count	symbol
56.0	collector
48.7	gibson
8.7	william
8.4	neuromancer
8.2	book
6.2	buy
4.8	novel
4.5	active
4.0	wait
4.0	request
4.0	eve
3.9	science
3.7	fiction
3.7	award
3.6	recognition
3.5	pattern

high.)

After filtering out stop-words, the average number of distinct symbols in the snippet-based representation of a keyword for 100 snippets is, in a sample of ten thousand representations, 710. Table 1 has the top fifteen symbols in the representation for “java”, which has 512 distinct symbols and a total of 811 symbols.

To get the vocabulary vector for a web-page we start by taking the text in the web-page and breaking it up into a bag-of-words. Unfortunately, many web pages have relatively little text. They might be succinct, or they might be stubs, or they might be nexuses linking to content but offering little direct content themselves. Low vocabulary counts are, with this classification method, likely to lead to poor accuracy.

We solve this problem and expand the vocabulary associated with a web page by recursively downloading the pages to which the base result page links, up to a given maximum depth (in this case, 3), and provided that the links are on the same host as the original link.

The vocabulary vector for each page so spidered is normalized so that its magnitude is constant. Also, each page is assigned a weight equal to $\frac{1}{2^n}$ where n is its distance in links from the root page. Finally, since obtaining the html for web pages is relatively costly (taking up to a few seconds per page) we limit the number of pages required by setting a maximum depth and, for web pages having more than ten links, choosing ten links at random.

In practice, this produces a characteristic vocabulary vector with on the order of four thousand distinct terms (after stop words and extraneous matter like java-script code have been discarded), which provides sufficient contextual discernment for our purposes.

It is easy to imagine this approach to modelling the semantics of web-pages failing. Web-pages often link to pages that are only peripherally relevant, or contain text that is digressive or irrelevant. Nevertheless, empirically (see below) this method works well.

One of the queries discussed below is *“gisbon context:neuromancer”* - one of the most relevant result pages for this query is *http://www.williamgibsonbooks.com/*, a part of whose characterization is in table 2

We compare semantic models using a Naive Bayes classifier. We approximated lexical prior probabilities by reference tothe British National Corpus (Leech et al 2001), which lists every word (and its frequency) in a large, heterogenous cross section of English documents, along with its frequency.

The score given in the tables below is the natural log probability of the normalized vocabulary vector of the web page being generated by the normalized vocabulary vector of the contextual query, divided by the number of symbols in the latter vector.

2 EXPERIMENTS

“Gibson” can refer to many things, including science fiction author William Gibson (whose first novel was “Neuromancer”), the Gibson Guitar Corporation (who also make basses), and actor Mel Gibson (who was in the move “Lethal Weapon”.) “Gibson”’s polysemy means that, for each of the intended interpretations of the term, there will be a large number of

Table 3: Top links for “gibson ctxt(neuromancer)”. Total good links = 7.

score	rank		url
0.79	5	X	http://www.williamgibsonbooks.com/
0.73	31	X	http://www.antonraubenweiss.com/gibson/
0.42	25	X	http://www.georgetown.edu/irvinemj/technoculture/pomosf.ht
0.31	94	X	http://www.ibiblio.org/cmc/mag/1995/sep/doherty.html
0.23	9	X	http://en.wikipedia.org/wiki/William_Gibson_(novelist)

irrelevant search results.

FIX ME

We put the system to the test on set of ambiguous queries - “gibson”, “fencing” and “web spider”. We use the contextual queries (“neuromancer”, “lethal weapon”, “acoustic bass”), (“immigration”[1], “less than zero”[2]), and (“jumping spider”, “pageranl”), respectively. We construct a contextual semantic query for each of these and evaluate the relevance of the results generated. Recall that one of the essential advantages of this method is that the relevant pages need not actually contain the contextual query.

Why not discard all the apparatus associated with the query “gibson ctxt(neuromancer)” and just use the query “gibson neuromancer”? The answer is that the latter query will give us pages about the book *Neuromancer* but not about William Gibson and his work in general. When one wants contextual but not extremely narrow focus, the *context* operator is useful.

Our test queries were as follows:

1. gibson ctxt(neuromancer)
2. gibson ctxt(acoustic bass)
3. gibson ctxt(lethal weapon)
4. fencing ctxt(foil)
5. fencing ctxt(agriculture)
6. web spider ctxt(jumping spider)
7. web spider ctxt(pagerank)

We evaluate the system’s precision and accuracy (in the top 10 slots in the filtered search.) Table 10 and table 11 summarize the system’s results. Table 10 shows accuracy and precision over the top five highest scoring web-pages of the first hundred served by google for the root query, and table 11 does the same for the top 20 pages.

[1]In the early twentieth century many immigrants to the US passed through Ellis Island.

[2]Brett Easton Ellis is a well-known author, one of whose novels is called *Less Than Zero.*

Table 4: Top links for “gibson ctx(lethal weapon)”. Total good links = 5.

score	rank		url
0.48	81	X	http://www.the-movie-times.com/thrsdir/actors/melgibson.ht
0.18	65	X	http://www.starpulse.com/Actors/Gibson,_Mel/
0.16	13	X	http://www.imdb.com/name/nm0000154/
0.14	92	X	http://www.rottentomatoes.com/p/mel_gibson/
0.12	74		http://deb.org/

Table 5: Top links for “gibson ctx(guiatar)”. Total good links = 11.

score	rank		url
0.85	11	X	http://www.zzounds.com/cat–Gibson–3549
0.84	99	X	http://www.12fret.com/retail/ggibsel.htm
0.78	12	X	http://www.zzounds.com/cat–Gibson-Electric-Guitars–3102
0.69	33	X	http://www.samedaymusic.com/browse–Gibson–3549
0.54	34	X	http://www.samedaymusic.com/browse–Gibson-Electric-Guitar

As the tables show, the algorithm does well. The worst performing query is *fencing ctxt(agriculture)*. In the top 100 Google results for the query *fencing*, there are only seven relevant to agricultural fencing. When sorted by relevance to *ctxt(agriculture)*, the top three ranked sites are relevant, but after this the results deteriorate. This is because there is only a weak connection between the notion of agriculture and fencing as it is used in an agricultural context.

Tables 3, 4, 5, 6 and 8 present the top 20 score-sorted web-pages, of the top 100 returned by Google, for their respective queries. A URL has an *X* to its left if that URL is relevant to the contextual query. The correlation between relevance score and actual relevance (as judged by a human, is very high. The top four results are invariably relevant to the contextual query. The drop off in quality of result appears to take place near, if not precisely at, the inflection point on the sorted relevance scores. The primary apparent fault is that there are sometimes relevant web-pages separated from the cluster of relevant pages at the top of the rankings by a set of irrelevant pages.

Table 6: Top links for “fencing ctxt(foil)”. Total good links = 39.

score	rank		url
0.89	68	X	http://www.maryland-fencing.org /links.htm
0.78	98	X	http://en.wikipedia.org/wiki/Gi orgio_Santelli
0.71	53	X	http://www.mtsu.edu/ fencing/eq uipment.html
0.67	41	X	http://www.brown.edu/Athletics/ Fencing/links.html
0.66	28	X	http://www.va-usfa.org/etc/supp liers.html

Table 7: Top links for “fencing ctxt(agriculture)”. Total good links = 7.

score	rank		url
0.66	89	X	http://www.sheepandgoat.com/fen cing.html
0.14	84	X	http://www.agry.purdue.edu/ext/ forages/rotational/fencing/
-0.01	17	X	http://www.foothill.net/ ringra m/fenceopt.htm
-0.10	20		http://www.ahfi.org/
-0.10	38		http://www.latourdulac.com/fenc ing/

Table 8: Top links for “web spider ctxt(jumping spider)”. Total good links = 39.

score	rank		url
1.52	71	X	http://www.cirrusimage.com/spid er.htm
1.06	45	X	http://www.uky.edu/Ag/CritterFi les/casefile/spiders/fishin
0.99	32	X	http://www.fi.edu/qa99/spotligh t5/index.html
0.94	70	X	http://www.cirrusimage.com/spid er_nursery_web.htm
0.93	7	X	http://www.xs4all.nl/ ednieuw/S piders/Info/Construction_of

Table 9: Top links for “web spider ctxt(pagerank)”. Total good links = 59.

score	rank		url
0.50	92	X	http://www.newfreedownloads.com /Web-Authoring/Site-Managem
0.43	67	X	http://software.ivertech.com/Si teScan-WebSpiderLinkChecker
0.36	59	X	http://www.tomdownload.com/web_ authoring/site_management/s
0.35	15	X	http://www.searchtools.com/tool s/ows.html
0.28	14	X	http://www.searchtools.com/robo ts/robot-code.html

Table 10: Results summary for contextual search for top 5 most relevant web-pages.

query	acc.	prec.	total
gibson ctxt(lethal weapon)	0.80	0.80	6
gibson ctxt(neuromancer)	1.00	1.00	8
gibson ctxt(acoustic bass)	1.00	1.00	12
fencing ctxt(foil)	1.00	1.00	40
fencing ctxt(agriculture)	0.60	0.60	8
web spider ctxt(pagerank)	0.80	0.80	59
web spider ctxt(jumping spider)	1.00	1.00	40

Table 11: Results summary for contextual search for top 20 most relevant web-pages.

query	acc.	prec.	total
gibson ctxt(lethal weapon)	0.83	0.25	6
gibson ctxt(neuromancer)	0.88	0.35	8
gibson ctxt(acoustic bass)	0.92	0.55	12
fencing ctxt(foil)	1.00	1.00	40
fencing ctxt(agriculture)	0.62	0.25	8
web spider ctxt(jumping spider)	1.00	1.00	40
web spider ctxt(pagerank)	0.85	0.85	59

3 RELATED WORK

The literature has many approaches to search query disambiguation. (Allan and Raghavan 02) describes an approach in which search queries are clarified by means of automatically generated, corpus-derived questions intended to identify the relevant aspect of the initial query. (Burton-Jones et al 03) and (Storey et al 06) describes a system that uses structured semantic information, in the form of WordNet or other manually constructed ontologies, to automatically refine search queries. (Sanderson and Lawrie 00) describes a method for disambiguating queries by providing a topic hierarchy for users to negotiate. The HiB system (Bruza and Dennis 97) offers query refinement by means presenting the user with corpus-derived suggestions for expansion and contraction of the scope of the query. (Shen et al) describes an approach to classifying queries in an ontology. Given a query, the system passes that query on to various search engines - its primary source of data are the ODP[3] classifications of that query, but in the event that these are unavailable it uses a feature-set derived from the web-pages returned for that query by the search engines.

Query disambigutation is a form of sense disambiguation, the literature of which contains some corpus-derived techniques. (Niu et al 04) describes an approach to word sense disambiguation that is in some respects analogous to the work described here. Their work uses a similarity metric based on

[3]Open Directory Project.

LSA-derived representations of and shared vocabulary from the contexts surrounding the instances of an ambiguous keyword in a corpus - the senses of the word in question are then derived using unsupervised learning techniques. (schutze 98) presents a corpus-based approach to word-sense disambiguation. It is based on the idea that two instances of an ambiguous word have the same sense if they have second-order similarity - that is, if there is substantial overlap between the words that they co-occur with co-occur with.

The related work described in this section is mostly about providing methods guiding the user, with more or less automation, to the information he wants. This work is different in that it provides a powerful but intuitive language for the user to express what he wants.

4 CONCLUSIONS AND FUTURE WORK

The method described in this paper is simple but effective. This technique for non-lexical, semantic search works because of the existence of a very-large, multi-topical collection of corpora, in the form of the Internet, and a fast, efficient method for searching over it lexically (in this case, Google, though any search engine would do.) The key observation is that simple characterizations of the search-result pages for a query provide a reasonable characterization of that query's meaning that can be used to compute inter-document distances.

This paper used supervised learning techniques over queries and documents but these distance metrics could also be used with unsupervised clustering algorithms. There have been many papers about the shape of the Internet, with topologies based on connectivity (i.e., (Faloutsos et al 99)) - it would be interesting to use the technique described herein to derive the semantic topology of the Internet, though the bandwidth and processing power required to do such a project justice would be vast.

REFERENCES

James Allan and Hema Raghavan (2002). Using Part-of-Speech Patterns to Reduce Query Ambiguity. SIGIR '02, Tampere, Finland.

P.D. Bruza and S.Dennis. (1997) Query-reformulation on the internet: empirical data and the hyperindex search engine. In *Proceedings of the RIAO Conference: Intelligent Text and Image Handling*, pages 488-499, Montreal, Canada.

Andrew Burton-Jones, Veda C. Storey, Vijayan Sugumaran and Sandeep Purao. (2003) A Heuristic-Based Methodology for Semantic Augmentation of User Queries on the Web. International conference on conceptual modeling, ER'03, pp. 476-489,

Michalis Faloutsos, Petros Faloutsos, Christos Faloutsos (1999) On power-law relationships of the Internet topology. Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication.

Google, Inc. www.google.com

Cheng Niu, Wei Li, Rohini K. Srihari, Huifeng Li, Laurie Crist. (2004). Context Clustering for Word Sense Disambiguation Based on Modeling Pairwise Context Similarities. SENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Barcelona.

Geoffrey Leech, Paul Rayson, Andrew Wilson (2001). Word Frequencies in Written and Spoken English: based on the British National Corpus. Longman, London.

Overture, Inc. www.overture.com

M. Sanderson and D. Lawrie. (2000) Building, testing, and applying concept hierarchies. In W. Bruce Croft, editor, *Advances in Information Retrieval: Recent Research from the CIIR, W. Brude Croft, ed., Kluwer Academic Publishers*, chapter 9, pages 235-266. Kluwer Academic Press, 2000.

Schutze, Hinrich. (1998) Automatic Word Sense Discrimination. Computational Linguistics. 24:1, 97-123.

Dou Shen, Rong Pan, Jian-Tao Sun, Jeffrey Junfeng Pan, Kangheng Wu, Jie Yin, Qiang Yang. (To appear.) Query Enrichment for Web-query Classification. ACM Transactions on Information Systems

Veda C. Storey, Andrew Burton-Jones, Vijayan Sugumaran, Sandeep Purao. (Preprint, submitted to Information Systems Review.) Making the Web More Semantic: A Methodology for Context-Aware Query Processing.

Yahoo, Inc. www.yahoo.com