

Using common hypertext links to identify the best phrasal description of target web documents

Einat Amitay
einat@mri.mq.edu.au

*Language Technology Group
Microsoft Research Institute
School of MPCE
Macquarie University
NSW 2109
Australia*

*CSIRO
Mathematical and Information Sciences
Locked Bag 17
North Ryde, NSW 2113
Australia*

Abstract

This paper describes previous work which studied and compared the distribution of words in web documents with the distribution of words in "normal" flat texts. Based on the findings from this study it is suggested that the traditional IR techniques cannot be used for web search purposes the same way they are used for "normal" text collections, e.g. news articles. Then, based on these same findings, I will describe a new document description model which exploits valuable anchor text information provided on the web that is ignored by the traditional techniques.

The problem

Amitay (1997) has found, through a corpus analysis of a 1000 web pages that the lexical distribution in documents which were written especially for the web (home pages), is significantly different than the lexical distribution observed in a corpus of "normal" English language (the British National Corpus - 100,000,000 words). For example, in the web documents collection there were some HTML files which contained no verb or determiner (i.e. "the", "a", etc.) although there are more than 60 words in them (excluding the HTML tags). While the word "the" comprised around 3% of the whole web collection, in the English language collection (BNC) it comprises about 7%.

This study also found that on the web there is a convention with which people write their documents: there is a certain number of words used to describe other target documents in the anchor text (i.e. `text`), and there is a linguistic convention in "highlighting" these words in the phrase, sentence or list.

Previous work and solutions

This section describes previous suggested solutions to the problem of finding information on the web, taking into account its structure and the additional meta-data provided by web authors. The solutions are presented in a chronological order and it is interesting to note that, through time, solutions rely more and more on the information provided by the authors of the web pages (e.g. link structure, anchor text, etc.). Since this paper suggests using the information embedded in the anchors and the link structure, the studies described below were chosen in order to show past trends in using such information.

Frei and Stieger (1992) describe a way for using the semantic content of hypertext links for retrieval. They present an indexing algorithm which makes use of the document's text and link content. The content of the link is marked as being "referential" or "semantic". Semantic links are further marked for textual content and pointing/node relations.

McBryan (1994) suggests that searches can be performed on titles, reference hypertext, or within components of URL name strings. In his system he indexes each URL with its anchor and title of page plus

the title of the target page "[in order] to maximise the available contextual information".

Weiss et al. (1996) combine link structure and textual information search techniques, by Performing TF-IDF on single documents and then on a higher level collection and so on, generating "content labels" for link clusters. The search is performed on the labels of the clusters and then - gradually - on a document level. Synonymous information is added to the labelling or, when needed, some of the more frequent terms in the cluster labelling are filtered. In this implementation documents are considered similar if they share similar links or are linked by similar documents.

Harmandas et al. (1997) combine text content and the nature of the web structure to find query terms in relation with images. They found that the texts immediately surrounding the images and the pages pointing at the pages where the image appeared on were much more useful for the search than the text which was not surrounding the image but appeared in the same node.

Kleinberg (1998) presents examples where the words in the query were not found in some of the most relevant documents: "web browsers" not found in Netscape's homepage, "Gates" not found in Microsoft homepage, etc. The author makes use of the directionality of links - i.e. pointers and targets, to compute hubs and authorities. He also points out that many of the authorities contain very little text which present difficulties using text-based search techniques (TF-IDF). The approach he suggests is based on co-occurrence of links and links' directionality in large collections of links (online indexes and lists). Since the problem of query terms does exist he suggests searching in the spatial "area" of the hubs and authorities (some kind of semantic network from the connected pages). He then uses the frequency of the terms in the defined "area". It is observed that hub pages tend to have higher occurrence of terms associated with the topic of the defined "area". This might be a good indication that the anchors contain better descriptions of the "area" or target page since hubs are large collections of links.

Brin and Page (1998), and Page et al. (in progress) rely on link structure and link text to provide information for making relevance judgements and quality filtering. They associate the text of a link with the page that the link appears on, as well as with the page the link points at. They also claim that anchors often provide more accurate descriptions of web pages than the pages themselves and that anchors may exist for documents which cannot be indexed by a text-based search engine, such as images, programs, and databases. Anchor propagation is used mostly because of the notion that anchor text can help provide better quality results.

Chakrabarti et al. (1998) introduce a notion that the text around href links pointing to a page is descriptive of the content of this page. If text descriptive of a topic occurs in the text around an href pointing to a page from a good hub, it reinforces the belief that this page is an authority on the topic. Assuming that the string `` would typically co-occur with the text Yahoo in close proximity, they study - on a test set of over 5000 web pages drawn from the web - the distance to the nearest occurrence of "Yahoo" around all href's to <http://www.yahoo.com> in these pages. Their results suggest that most occurrences are within 50 bytes of the href. Qualitatively similar experiments with href's other than Yahoo suggested similar results. It seems that for the output of their system the wording of the anchors is more significant in the relevance-to-topic ranking than the source page these anchors appear on.

A new model for presenting search results by using the data provided by web users

The work described below is part of my PhD and it outlines a model for the solution suggested. The results are not yet publishable and therefore this section will only describe the idea itself.

Motivation

Following Amitay's (1997) findings about the consistent linguistic behaviour with which web authors write their pages, in particular the wording of the anchor text, this model makes use of the descriptions provided by those authors for creating a collaborative description tool for web pages.

The web, as the name suggests, is a large collection of documents "glued" together in different places. The "glue" in this case is hypertext links. Links consist of hidden-to-the-user URL addresses and visual cues

such as phrases or images which are physically distinct from the surrounding environment.

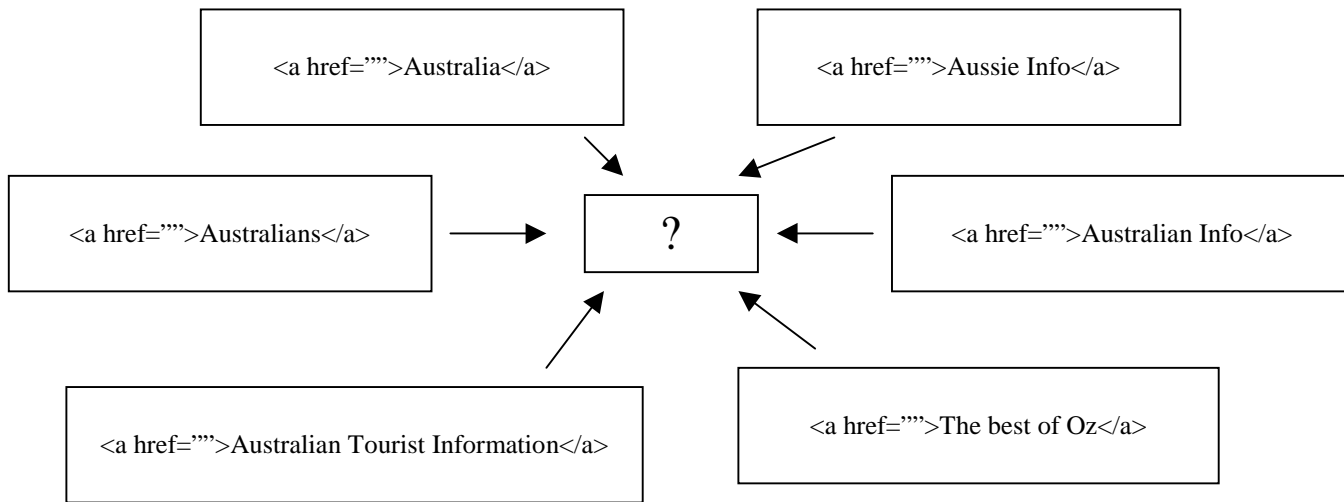
Since on the web many documents are pointed at by more than one link, it might be useful to find similarities between pointing anchors. Such a comparison would allow us to formulate a relation between links and their target document. By formulating such a relationship we might be able to improve automatic "understanding" of web documents, as well as to the understanding of the way people construct web documents.

Anchor wording provides an insight into the way people prefer to name their links. By providing few words, a short phrase or sentence they provide hints as to what might be found on the target page. This short description of the target page could prove very helpful if we are able to compare it with other anchors pointing at the same document. Extracting the anchors to a web document from the documents pointing at it could provide valuable information as to the content of the target web document.

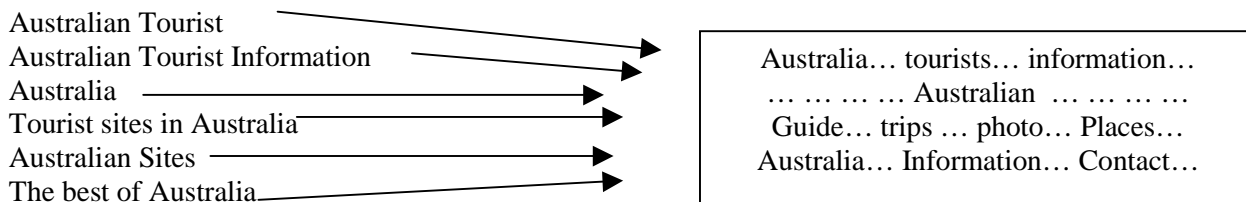
Method

As mentioned earlier, this is only an outline of the idea and I hope to be able to publish some results in the next couple of months. The stages of the implementation of the algorithm are listed below and the system described would be placed on top of an existing search engine which has an available "map" of the web (Such as the one described by Brin and Page (1998)). This system would present the results from the engine by returning the found documents together with a short description generated from the anchors pointing at it.

Stage 1: Collecting lists of URLs plus anchors which point to the same target page. This collection should be mapped for documents' spatial relations.



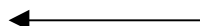
Stage 2: Comparing the wording of the anchors which point at the same target for similarities, phrases similarity, length, overlapping information vs. non-overlapping information provided by different anchors, "grammaticality" of syntactic constructions, etc.



Stage 3: Looking at the significance measure of the words chosen as anchors in the target document using the burstiness algorithm (Richmond, Smith and Amitay, 1997), TF-IDF (Salton and McGill, 1984; Salton and Buckley, 1988), "top-page-information" and meta-information.

Stage 4: Reversing the process to find which of the anchors predicted most of the topical keywords detected by the IR algorithms and the meta-data in order to be able to find the best description. This also involves "gluing" two anchors together, creating longer and shorter descriptions, etc.

Australian Tourist
Australian Tourist Information
Australia
Tourist sites in Australia
Australian Sites
The best of Australia



Australia... tourists... information... Australian Guide... trips ... photo... Places... Australia... Information... Contact...
--

Conclusion

The idea and algorithm described above is a follow-up on a work which studied in detail the way people write web documents (Amitay, 1997). It is also a continuation of the research trend which makes use of direct information provided by people writing for the web. Since web authors dedicate hours and days of their time for presenting and tailoring their documents it is about time that this information will not be ignored by the term statistics....

References

Amitay E. (1997). Hypertext: The Importance of being Different. MSc thesis, Centre for Cognitive Science, The University of Edinburgh, Scotland. Also a Technical Report - No. HCRC/RP-94.
<http://www.hcrc.ed.ac.uk/publications/rp-94.ps.gz>

Brin S. and Page L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. Proceedings of the Seventh World Wide Web Conference (WWW7), Brisbane, also in a special issue of the journal Computer Networks and ISDN Systems, Volume 30, issues 1-7.
<http://google.stanford.edu/~backrub/google.html>

Chakrabarti S., Dom B., Gibson D., Kleinberg J., Raghavan P., Rajagopalan S. (1998) Automatic resource list compilation by analysing hyperlink structure and associated text. Proceedings of the 7th International World Wide Web Conference, Brisbane, also in a special issue of the journal Computer Networks and ISDN Systems, Volume 30, issues 1-7.
<http://simon.cs.cornell.edu/home/kleinber/www98/438.html>

Frei H.P., Stieger D. (1992). Making Use of Hypertext Links When Retrieving Information. Proceedings of the 4th ACM Conference on Hypertext ECHT '92, ACM Press, New York, pp. 102-111.
<http://www.acm.org/pubs/citations/proceedings/hypertext/168466/p102-frei>

Harmandas V., Sanderson M., and Dunlop M.D. (1997). Image retrieval by hypertext links. Proceedings of SIGIR-97.
<http://www.dcs.gla.ac.uk/~mark/publications/harmandetal.pdf>

Kleinberg J. (1998). Authoritative Sources in a Hyperlinked Environment, Proceedings of the ACM-SIAM Symposium on Discrete Algorithms. Also appears as IBM Research Report RJ 10076, May 1997.
<http://simon.cs.cornell.edu/home/kleinber/auth.ps>

McBryan O.A. (1994). GENVL and WWW: Tools for Taming the Web. First International Conference on the World Wide Web. CERN, Geneva
<http://www.cs.colorado.edu/home/mcbryan/mypapers/www94.ps>

Page L., Brin S., Motwani R., and Winograd T. (in progress). The PageRank Citation Ranking: Bringing Order to the Web.

<http://google.stanford.edu/~backrub/pageranksub.ps>

Richmond K., Smith A., and Amitay E. (1997). Detecting Subject Boundaries within Text: A Language Independent Approach. In Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing, Brown University, Providence, RI, USA.

http://www.cogsci.ed.ac.uk/~einat/EMNLP2_code

Salton G. and Buckley C., (1988). Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*, 24(5):513-523.

Salton G. and McGill M.J (1984). *Introduction to Modern Information Retrieval*. McGraw-Hill:NY.

Weiss R., Velez B., Sheldon M.A, Manprempre C., Szilagyi P., Duda A., and Gifford D.K. (1996). HyPursuit: A Hierarchical Network Search Engine that Exploits Content-Link Hypertext Clustering. Proceedings of the Seventh ACM Conference on Hypertext, Washington, DC.

<http://www.psrg.lcs.mit.edu/ftplib/papers/hypertext96.ps>