Exploiting Anchor Text as a Lexical Resource

Peter Anick

Yahoo!
Sunnyvale, CA
panick@yahoo-inc.com

Abstract

Anchor texts, the strings associated with hyperlinks on a web page, are currently employed to express millions of referrals to sites and topics on the world wide web. We consider how these strings might be exploited as a lexical resource, particularly when viewed from the perspective of their target documents rather than their sources. We find that for many target pages, incoming anchors form a miniature corpus of reference expressions whose properties with relation both to other target sites and to each other can be put to use for mining lexical information.

Introduction

With a corpus of over 4 billion pages, the world wide web has become a rich source of textual data from which to build dictionaries of concepts and named entities. The fact that it is a hypertext medium offers opportunities to the data miner beyond those techniques developed for entity extraction from raw text (as in Riloff, 1996; Mikheev et al, 1999) . Chen et al (2003), for example, capitalized on the way that textual links are employed to structure web sites' subject matter in order to construct a web thesaurus. Lu et al (2001) extracted bilingual translations based on the co-occurrence of Chinese and English inlinks to the same target pages. Sundaresan and Yi (2000) mined the web for name-acronym pairs using rules that sometimes straddled raw text and markup language.

Anchor text, the text explicitly associated with a hyperlink on a web page, often serves to provide a succinct descriptor for the target document. This property has been extensively exploited by web search engines and has led to major improvements in relevance ranking, especially for "navigational" queries, whose intended target is the web site of a named entity (Brin and Page, 1998; Craswell et al, 2001). The success of utilizing anchor text for named entity and topical searches suggests that anchor text would also serve well as a lexical resource for applications such as dictionary construction and named entity extraction.

Since every hypertext link has both a source and target document, there are two distinct perspectives from which to analyze anchor texts. Viewing a link from the perspective of the source document allows one to examine it within the textual context in which it appears. This is the view from which most research on entity extraction has been carried out, and the existence of hypertext mark-up essentially serves as one further clue that a region of text is deserving of special scrutiny. The perspective of the target document, however, provides an opportunity to analyze anchor texts in a different manner, since all the incoming text to a specific target can be treated collectively as a miniature corpus of reference expressions, whose properties with relation to target sites and to each other can be inspected. This is the approach we pursue here.

The paper consists of two parts. In the first section, we survey the general properties of anchor text data when viewed from the perspective of target documents. In the second, we delve deeper into an area for which anchor text appears to be highly suited as a lexical resource – for the capture and analysis of proper names and their variants.

A geological survey

Our raw data was produced by a web crawl (of roughly 2 billion pages) by the spider employed by the AltaVista search engine. All anchor texts on crawled pages which referred to a page *external* to the source site were saved, along with their source and target URLs. For each target URL, the set of incoming anchors was accumulated to create a database of records of the form*: target_url anchor_text count.* Counts were summed across normalized strings, which were case folded and had some punctuation (such as surrounding quotes or brackets) removed. Internal punctuation, such as parenthesized substrings, commas, periods, and apostrophes, were retained. For the purposes of this study, we selected an arbitrary subset of this corpus, consisting of roughly 1.5 million URLs.

Web addresses (uniform resource locators) contain a host name, optional directory path, and file name. We will define the depth, or level of the URL to be the length of the directory path. For most home pages, the directory path is empty and longer directory paths indicate greater depth within a web site. As a result, one might expect the nature of incoming anchor text to differ from level to level. Tables 1 and 2 show how the total number of inlinks and the number of different inlinks changes from level 0 to level 5 URLs. Keeping in mind that we are only considering inlinks from sites other than the target site, the data show that higher level URLs tend to have both more inlinks and more inlink diversity. That is to say, the sites

are not only more highly referenced but they are referenced using multiple textual descriptions. [1]

	URL depth					
# diff. inlinks	0	1	2	3	4	5
1	69	71	78	82	88	87
2	17	11	13	10	9	7
3	6	7	2	4	1	4
>3	8	11	7	4	2	2

Table 1: Percent of URLs with 1, 2, 3, >3 different inlinks at URL depth of 0 to 5.

	URL depth					
# inlinks	0	1	2	3	4	5
1	45	33	52	60	69	62
2	12	15	17	18	12	17
3	11	9	8	6	6	7
4-10	19	33	17	12	7	10
>10	13	14	6	4	4	4

Table 2: Percent of URLs with 1, 2, 3, 4-10 ,>10 inlinks at URL depth of 0 to 5.

For each URL, we define as its default "top anchor" the anchor text string with the highest inlink frequency, not counting anchors which are essentially a representation of the URL itself (e.g., "www.altavista.com"). Random sampling of top anchors at each URL depth reveals that the mix of lexical types varies considerably by level. Of the top anchors for level 0, over 50% are entity names (e.g., *park shore bmw*, *farmington public library*), 25% are nominal concepts (*outboard motors, audio video*), and 8% personal names. Level 5 URLs, by contrast, are dominated by "headline-like" anchors – longer, more syntactically rich specifications that would be appropriate for the header of a news article or chapter on some topic. For example:

- about building a family tree for kids
- new voice for teachers
- motorola plans $1.9 billion investment

In level 5, the percentage of named entity anchors drops to 28%. For the most part, the entities found at this level are region names, such as "france" or "united kingdom" rather than the kinds of organization names directly associated with specific web sites that dominate the higher levels.

[1] The tables show a slight increase in inlink diversity and count at depth 1. While this requires further analysis, we suspect that some of the increase is due to a higher degree of spammed sites at this level, i.e., sites for which linkage has been artificially manipulated.

Anatomy of a target inlink set

The preponderance of named entity targets in level 0, along with the large number of targets with diverse inlink sets makes this corpus particularly attractive for both lexicographic research into and automatic acquisition of proper names, variants, and segmentation behavior. Sorting the inlinks by frequency for each target page allows us to compare lower frequency anchor strings to the "top anchor" string. Specifically, we can classify each lower frequency anchor according to its superficial lexical relationship to the top anchor as follows:

SS: specialization – a string which has the top anchor as a substring
SU: substring – a string which is a substring of the top anchor
ST: a string which is neither an SS or SU but shares some term(s) in common with the top anchor
AC: a possible acronym for the top anchor
UR: a likely URL name
UN: a string not related to the top anchor in any of the above ways

The example below shows the count, anchor text, and classification of the anchor into one of these categories.

45	desert tortoise preserve committee	TA
10	[the] desert tortoise preserve committee	SS
5	www.tortoise-tracks.org	UR
3	desert tortoise preserve committee[, inc]	SS
2	desert tortoise preserve [committee]	SU
2	desert tortoise preservation committee	ST
2	desert tortoise [preserve committee]	SU
1	tortoise tracks	ST
1	dtpc	AC
1	desert turtle preserve committee	ST
1	desert tortoise preserve committee[, the]	SS
1	desert tortoise natural area	ST
1	desert tortiose preserve committee	ST

The anchors include name variants, the URL, topics that the name relates to, an acronym, and a misspelling. For SS anchors, we extract the prefix and suffix strings, that is, the portions of the SS string that extend the top anchor to the left or right (shown above by adding bracketing). As the top anchor often contains the most official name for the organization, SS prefixes and suffixes tend to contain optional name qualifiers, as well as "noise" words such as "click here for" and "site". For the SU anchors, we extract those portions of the top anchor name that would have to be "dropped" in order to form the SU string (shown in brackets on subsequent lines). These tend to be segments of the official name that may be elided, such as "association", "magazine", and "university". By capturing the most common prefixes and suffixes found for SU and SS anchors, we can assemble a list of English

organization type identifiers.[2] To help separate out noise terms (heads such as "website" or "home") from content terms in these lists, we computed the ratio of suffixes found in SS strings to those found for the SU anchors. Since noise terms are more likely to be added to the top anchor name than removed from it, sorting by this ratio creates an ordered list with most noise words appearing at the top and stronger content terms appearing at the bottom, as in the following examples (showing ratio, SS count, SU count and anchor, computed over a subset of level 0 target pages with inlink count > 9 and inlink diversity > 3):

31.5	63	2	official web site
31	93	3	official site
25	25	1	listings
23	23	1	's own web site
…			
3	9	3	germany
3	9	3	daily
3	9	3	community
…			
0.3	9	26	society
0.3	9	26	school district
0.3	6	18	news
0.3	35	107	association

Mining named entities

Data mining from textual sources typically employs both internal and external evidence for the identification and categorization of terms (McDonald, 1993). Internal evidence refers to evidence within the term itself, such as head words like "school" and "association" that associate a proper name with a semantic category. External evidence comes from surrounding context, such as verbs and appositives. Working with disembodied anchor text removes many opportunities for exploiting such external clues. On the other hand, anchor text by its very nature comprises a relatively concise compendium of the entities and topics of interest on the web, and the term delimitation problem is simplified by the fact that many anchor texts are already self-contained multiword units. Furthermore, as noted above, frequencies and surface string relationships among anchors associated with the same target can be exploited to derive head terms and lexical sub-contexts within these referencial expressions. Similarly, one can draw inferences from other sorts of external data, such as the number of targets an anchor text is associated with, the URL depth of the target, etc. In this section, we briefly describe work in progress to capitalize on such clues for term extraction from anchor text corpora.

[2] The English language tends to dominate within anchor text for organizations on the web. However, by partitioning anchors according to the language of the source document, it should be possible to carry out language-specific analyses of the same kind described here.

Top anchor frequency analysis

Most organizations have a single web site (modulo regional offices). It is reasonable to assume that top anchors which capture legitimate organization names should have relatively few different target URLs. Therefore, by sorting our default top anchors by target count, we can capture those anchor strings which are more likely to be topics or qualifiers rather than entity names. Examples of anchors found at various target count levels within our level 0 URL depth sample are shown below.

1264 home
143 click to enter
62 new york
61 flowers
19 auto insurance quotes
7 linux
3 france telecom
1 yukon lions club

Such count evidence can be used to disqualify as entity names those anchors that for other reasons have received the highest inlink frequency for their respective target sites. For example, the highest frequency inlink for the "new york comedy club" site is the anchor "new york". Disqualifying it as top anchor allows "new york comedy club" to rise to the top anchor spot.

Head terms

From the ratio sorted list derived above from SU and SS segments, we extracted 154 organization "head" terms, to use as lexical evidence for the presence of a named entity. Candidate entity names were drawn from the top anchors of sites that passed a crude threshold for "importance" based on their total number of inlinks and the number of different inlinks to the site. From a pool of 1.5 million sites, we found 201,019 sites with at least 10 total inlinks and 4 different inlinks. Matching the top anchors of these sites against the list of head terms (allowing heads to appear either at the end of the string or just preceding the prepositions "of" or "for") yielded 42,330 named entities, accounting for 21% of the sites.

Acronyms

Acronyms tend to appear in two forms within anchor texts. They may be placed within parentheses after the full name, or they may appear as independent anchor texts, usually with lower link frequency than their full name counterpart. In either case, because of the strong implied relationships among terms associated with the same target URL, one can apply relatively loose matching criteria to associate potential acronyms with their full names. Counting numeric sequences as a single unit, we look for acronyms that contain > 50% of the start letters of the corresponding non-noise anchor components and whose length does not exceed the number of non-noise components by > 2 units. This heuristic covers those

common cases in which acronyms contain more than one (not necessarily adjacent) letter from the same source word or do not include initials for all content words, as in the following examples:

executive education network (exen)
wildlife care international (wlci)
the 2003 conference on multimedia computing and networking (mmcn2003)

A special case are those variants which contain a mix of acronym and full words. These can be detected by first scanning for substring matches and then applying the acronym matcher on the remainder of each candidate name, as in

international federation of business and professional women (bpw international)

From our test corpus of 201,019 "important" level 0 sites, we extracted nearly 13 thousand acronyms appearing as separate anchors within a target's inlink set and over 28 thousand appearing in parentheses after the full name.

Anchor text segmentation

The list of SS and SU strings derived from the substring analysis of anchors within each target page's inlinks can serve as a dictionary of common segments for parsing anchor texts in general. The segments captured in this way include not only noise phrases and entity name head terms but also a number of other phrases which can be exploited for segmenting multi-concept anchors and, in some cases, for predicting the semantic types of the components of multi-concept anchors. For example, locations can often be found in such contexts as

city of, embassy of, hotels in, travel to, buy home in, …,

celebrity names are typically to be found with

fan site, fan club, fan page, fansite, fans website,…,

and products appear in the context of

buy, shop for, …

Thus, while anchor texts divorced from their source pages lose most of the broader textual context that might provide further clues about their semantic classes, there are nonetheless some category clue terms that conventionally appear within the bounds of the anchor text strings themselves. The extent and reliability of such self-contained phrasal contexts are a subject of further investigation.

Conclusions and future research

As a collection of referring expressions to web sites and topics, anchor text holds the promise of providing a concise lexical representation of web content at a fraction of the size of the full text of the indexable world wide web. We have investigated a number of properties of an anchor text corpus organized from the perspective of its target pages in order to assess its potential as such a lexical resource. We have found this perspective particularly useful for the analysis and extraction of named entities, variants, and acronyms.

As a next step, we plan to refine the techniques outlined here using a much larger sample, which should enable us to tune statistical parameters needed to improve the process of top anchor selection/qualification and anchor text segmentation. A second objective is to apply our techniques to anchor text corpora built from source pages in non-English languages. Finally, we plan to investigate the properties of internal anchor text, that is, anchors which refer to pages within the same web site as the source. For such anchors, however, it is likely that the perspective of the target page may be less informative than for external links and many of the properties noted here for external links may not apply.

References

Brin, S. and L. Page (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. In Proceedings of The Seventh International World Wide Web Conference, 1998.

Chen, Z., S. Liu, L. Wenyin, G. Pu and W. Ma (2003). Building a Web Thesaurus from Web Link Structure. In Proceedings of SIGIR'03, 48-55.

Craswell, N., D. Hawking and S. Robertson (2001). Effective Site Finding using Link Anchor Information. In Proceedings of SIGIR'01, 250-257.

Lu, W., H. Lee and L. Chien (2001). Anchor Text Mining for Translation Extraction of Query Terms. In Proceedings of SIGIR'01, 388-389.

David McDonald (1993). Internal and External Evidence in the Identification and Semantic Categorization of Proper Names. In Acquisition of Lexical Knowledge from Text: Proceedings of a Workshop Sponsored by the Special Interest Group on the Lexicon of the Association for Computational Linguistics (B. Boguraev and J Pustejovsky, eds.)

Riloff, E. (1996). Automatically Generating Extraction Patterns from Untagged Text. In Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96), 1044-1049.

Sundaresan, N. and J. Yi (2000) Mining the Web for Relations. Computer Networks, 33(1-6): 699-711.