Searching for Content-based Addresses on the World-Wide Web

Joel D. Martin
Interactive Information Group
National Research Council
Ottawa, ON Canada K1A 0R6
E-mail: joel@ai.iit.nrc.ca

Robert Holte
SITE, University of Ottawa
Ottawa, ON, Canada

E-mail: holte@site.uottawa.ca

ABSTRACT

This paper presents a method for constructing queries that are sufficient to retrieve a target web page. These queries can be thought of as content-based addresses for the target page and can have many potential uses.

KEYWORDS: distributed digital libraries, query, web, search engines, content-based addresses, dead links, QuerySearch

INTRODUCTION

In a global network of digital libraries or indeed any distributed digital library, address-based document identifiers can fail in many ways. Documents can easily be lost if any part of the address changes, i.e., if the document moves or the domain name changes or disappears. One solution to this problem is to introduce Universal Resource Names as persistent, globally unique identifiers just like the ISBN for printed books (1). A complementary solution that does not require any naming authority and that has additional benefits is to create content-based addresses.

The World Wide Web (2) is one model of a distributed, global, digital library and its Universal Resource Locators or URL's are the address-based document identifiers. URL's frequently become ineffective and cause web pages to be lost. A content-based alternative to a URL would be a list of key terms that could be used as a query to retrieve the target web page from a large search engine. We will call this query a *content-based address* or more specifically a *summary-query.*

A content-based address would not change when a target web page has moved. As long as the big search engines keep crawling and indexing, the content-based address will still work. Additionally, a content-based address, if properly constructed, could be used to find pages that had both moved and changed and may find other pages on the same topic.

Even if the search engine indexes have grown enough that the target web page is no longer at the top of the results list, it will still be there, just ranked lower than before.

The purpose of this paper is to demonstrate that two types of content-based addresses *can* be found for many web pages. a) Long, high precision queries can be found that will locate the target web page even if it has moved; and b) several short queries can be found that together permit locating the page after it has moved and been modified.

ARE SUMMARY-QUERIES POSSIBLE?

Consider a simplified web in which there are D = 1,000,000,000 web pages and that every web page contains w = 1000 distinct words chosen from the W = 200,000 possible words. Further, assume that all words are equally likely and appear independently so that if word A appears in a web page, that does not affect the probability that word B appears in the page.

In this simplified world, a query of m = 1 word would retrieve D * (w/W)^m = 5,000,000 pages. A conjunctive query of two words would retrieve 25,000 pages and a conjunctive query of three words would find 125 pages. If m = 5, the expected number of pages returned is less than 1. As a result, if we extract any five mostly independent words from the target web page, we will retrieve this page and very likely nothing else.

In the real world, there are correlations between words, requiring somewhat longer queries. However, on some search engines, queries can incorporate phrases, thereby allowing somewhat shorter queries.

FINDING CONTENT-BASED ADDRESSES

QuerySearch is a system designed to search for a query that results in one or more particular documents being retrieved. There is a toolbox of possible search heuristics that can be applied. Basically, an initial query is simplified or extended in order to find a query that does a better job of finding the target documents.

In this paper, we describe the application of QuerySearch to the problem of finding summary-queries on Alta Vista. AltaVista was chosen because of its large coverage of the web and the fact that it indexes a large percentage of the

words on every web page. QuerySearch has also been used with MG (3) and Infoseek.

In this application, an initial query is generated and then is simplified in many ways. The initial query was designed to be a long, high precision query that would be helpful for finding moved documents. The short, simplified queries could be used together to find documents that have been changed.

After any query is created or simplified, the modified query is submitted to the search engine to verify that the query is successful. A query was considered successful if the target web page appeared in the top 10 results from the search engine.

Overall process

Initial Query generation

1) Extract words and phrases from page.
2 Rank the terms by frequency and position.
3 Conjunctive query: *f* best terms and title terms.
4 Verify that page is retrieved in the first hit of the results. Otherwise, increase *f* and repeat 3.

Backwards Elimination

1) Remove a query from a query queue.
2 For every term, simplify it, submit the query
3 If in the top 10 results, put query on queue
4 If query cannot be simplified, save in results.
5 Repeat from beginning

Simplification:

1) changing a 'must' ('+') to a 'should'
2 splitting a phrase into parts
3 change a 'title' word to a word
4 drop a word from the query

FOR EXAMPLE:

When the above algorithm is applied to the preliminary program for Digital Libraries '97, an initial query of 9 terms (2 words, 7 phrases) is simplified to several queries of 2 terms.

URL:
http://www.sis.pitt.edu/~diglib97/PreliminaryProgram.htm

Initial Long Query:
'+chair +session +"digital library" +"digital libraries" +"coffee break" +"stanford university" +"carnegie mellon university" +"preliminary program" +"edie rasmussen" '

Four Simplified Queries:
+"stanford university" edie rasmussen
+"coffee break" preliminary edie
+"preliminary program" rasmussen
+"digital library" "preliminary program"

As of 13 April, 1998, the initial query above retrieves one web page, the target, while the simplified queries each retrieve more than 10,000 web pages with the target web page in position 2 - 7.

INITIAL RESULTS:

The QuerySearch system was applied to a set of "random" web pages, i.e., the list of 'new' web pages that were indexed on Yahoo April 6, 1998 (URL). To be used in the experiment, the web pages had to be on the web for at least one week, indexed by Alta Vista, and not be 'frameset' pages.

There were 756 remaining web pages. For each of these, the above procedure was run and the size and success of the initial query and best final queries were calculated. QuerySearch found successful queries for all 756 web pages.

	Avg Length in terms	target position
Unsimplified	8.6	1.24
Simplified (5 shortest)	2.34	6.54

DISCUSSION

It is possible, even in the real world of millions of web pages, to translate a web page into an "equivalent" short query. The query is equivalent to the web page in the sense that the query will retrieve the web page in the top 10 results using a major search engine. This is a form of content-based addressing.

These queries can help locate moved or changed pages. When the web page moves, the address still works. When web pages are added, only those that match the same query will be placed above the target web page in the results list.

QuerySearch actually collects *all* successful short queries. Any of these queries would be acceptable content-based addresses. If one fails, which would be the case for changed pages, another query could be used.

The queries found by QuerySearch can also be used to locate related pages. Summary-queries should contain words or terms that are most relevant to the meaning of the target page. Currently, QuerySearch uses a bias that attempts to mimic AltaVista's weighting scheme and takes frequency and position of occurrence into account.

A FINAL WORD

Content-based addresses will never replace URL's. However, they can help alleviate the problem of moved or changed pages in a distributed digital library. In addition, they show promise for finding related web pages.

ACKNOWLEDGMENTS

Thanks to Peter Turney for stimulating suggestions and discussions related to the above ideas.

REFERENCES

1.Karen Sollins & Larry Masinter. "Functional Requirements for Uniform Resource Names", Internet RFC 1737, (1994)

2. Tim Berners-Lee, Robert Cailliau, Jean-François Groff, Bernd Pollermann: World-Wide Web: The Information Universe. *Electronic Networking: Research, Applications and Policy* 1(2): 74-82 (1992)

3. T.C. Bell, A. Moffat, I.H. Witten, & J. Zobel, The MG retrieval system: compressing for space and speed, *Communications of the ACM*, *38*,41-42, (1995).