Assessing Intra and Extra Web-based Automatic Indexing Tools

Elaine G. Toms*
Faculty of Information Studies
University of Toronto
toms@fis.utoronto.ca

Toby Reid
Compaq Canada

[*person to whom all correspondence should be sent]

ABSTRACT

In this research we compared two types of indexing tools: one based on the content of a document, an intra-indexer (*Extractor*), and one based on a document's external representation – the nodes to which it is linked (*Topic*). This paper summarizes the results of a pilot study and thus must be cautiously interpreted.

INTRODUCTION

Many tools have been developed for the automatic indexing of documents. They result from a myriad of techniques such as neural networks, learning algorithms, term frequency and so on. The common thread among them is their use of the contents of a document as the source for generating indexing terms. With the creation of the WWW, the interconnected network of nodes has provided an additional option – the identification of content based on links (Kleinberg, 1998; Chakrabarti et al 1999, 2000; Rafei & Mendelzon, 2000). Analogically, this is a similar approach to citation analysis; it is based on the assumption that when two documents are linked, the content of the source is somewhat indicative of the content of the destination. Google is perhaps the best known of this class of tool.

In this research we compared these two approaches: a) an 'extra' indexer, one that indexes a document according to how it is represented external to the document, and b) an 'intra' indexer, one that indexes a document using the contents of that document. We selected two accessible tools that are representative of each type and compared their output. We hypothesized that an intra-document automatic indexer provides a better (i.e. more accurate and relevant) set of key-words/phrases than an extra-document automatic indexer.

METHODS

Indexing Tools

Topic (http://www.cs.toronto.edu/db/topic/) is an example of an 'extra' indexer (Rafiei & Mendelzon 2000) while *Extractor* v.6.1 (http://extractor.iit.nrc.ca/) is an example of an 'intra' indexer (Turney, 2000). *Topic* determines the topicality of a particular document, not by analyzing a document's textual patterns, but by determining what the document is reputed to be about. It algorithmically computes a document's topic reputation according to the reputations of external web pages that hyperlink to it, and, furthermore, of other web pages that link to those pointing-pages as a means of corroborating the calculated reputations. *Extractor* is a machine-trained system leveraging traditional IR term-level operations (i.e. a ten-step intra-document key-phrase extraction algorithm employing twelve pre-tuned parameters). Thus, *Extractor* automatically indexes a given document by algorithmically analyzing the textual patterns comprising it. The environment in which it operates is that of the document itself.

Both tools were chosen because of their ease of access; both are available on the WWW. *Topic* has not been rigorously evaluated (see Keast, Toms & Cherry, 2001 for one of the few user-based evaluations of *Topic*'s output). *Extractor* has been compared with four other indexing tools (Turney 1997) and is currently being used by other types of tools. Both appear to behave similarly: they accept a URL as input, and output a set of indexing words.

Site Selection

We randomly selected four private sector industry categories from the Business section of the Yahoo directory (http://dir.yahoo.com/Business_and_Economy/Business_to_Business/). The Business section was deliberately selected because of the expertise of team members. Within that section, we randomly selected three companies from each of the four industry categories. From each company web site, we used the site-level (i.e. homepage) URL, and a site-level-minus-one URL. This procedure resulted in 24 web documents (4 industries x 3 companies x 2 URLs per company). See Table 1 for a complete list.

Procedure

At the time of document selection, each web document was indexed twice – once using *Topic* and once using *Extractor*. Thus 48 index sets (24 web documents x 2 indexes per document) were obtained. For each of the 24 web documents in the test, a 30-term list of key-words/phrases was created. This was achieved by merging the output terms produced by *Extractor* and *Topic*. Duplicate terms were excluded. The resulting lists were randomized.

Copies of the 24 web documents and the corresponding indexing term list were given to a trained indexer for ranking and rating purposes. A human indexer's rating and ranking measures acted as a control for comparing the index results. The indexer was given a rating sheet for each and instructed to: rate on a scale of 1 to 5 the relevancy of each keyword to the topic of the page, and rank five keywords that best indicate what the page is about. No indication of the source of a keyword was provided. The results were initially input into a Microsoft Excel spreadsheet for subsequent data analysis using SPSS.

Yahoo Sub Category	URL
Office Supplies & Equipment	http://www.inkspot.com
	http://www.inkspot.com/faqs.htm
	http://www.pirongs.co.uk/
	http://www.pirongs.co.uk/trade.html
	http://www.filingsystems.com
	http://www.filingsystems.com/services/index.html
Cleaning / Services	http://www.environmentalcontrol.com
	http://www.environmentalcontrol.com/history.html
	http://www.dowservices.com
	http://www.dowservices.com/miss.html
	http://www.idsglobal.com
	http://www.idsglobal.com/Pack/Pkg_Nov_Member_info.htm
Business Opportunities	http://www.homeopportunities.com
	http://www.homeopportunities.com/security.htm
	http://www.clientconnectioninc.com
	http://www.clientconnectioninc.com/products.asp.html
	http://www.jinglebiz.com
	http://www.jinglebiz.com/webpq2.htm
Textiles / Manufacturers	http://www.camronnet.com
	http://www.camronnet.com/net.html
	http://www.elastix.com.au
	http://www.elastix.com.au/index16.htm
	http://www.rayonyarn.com
	http://www.rayonyan.com/texapps/

Table 1. Sites used in Test

RESULTS

Extractor generated keywords for 22 of the 24 sites while *Topic* could provide keywords for only 5 of the sites. Overall, the tools jointly indexed only four sites. As a result, we were unable to assess statistically significant differences between the two tools. Unindexable sites tended to be those in which words were embedded in graphics or were dynamically generated so that the indexing tools could not interpret the page. Topic's unusual behavior may also have been because the pages were so extensively linked by other groups that its intent could not be determined.

Extractor output from 0 to 30 keywords per site, averaging 17 keywords for each site. *Topic* output from one to nine keywords for the five sites that it indexed (averaging across all sites will produce a nonsensical number). Forty-seven per cent of *Extractor*'s index terms were in the top ratings (from 3 to 5). Seventy-nine per cent of *Topic*'s terms were given the same rating. Because results are based on only five sites, these results must be interpreted judiciously and cautiously. See Table 2 for the results for each tool.

URL	Extractor	Topic
http://www.inkspot.com	7%	22%
http://www.inkspot.com/faqs.htm	33%	No terms
http://www.pirongs.co.uk/	68%	No terms
http://www.pirongs.co.uk/trade.html	100%	No terms
http://www.filingsystems.com	58%	No terms
http://www.filingsystems.com/services/index.html	43%	No terms
http://www.environmentalcontrol.com	28%	No terms
http://www.environmentalcontrol.com/history.html	71%	No terms
http://www.dowservices.com	31%	No terms
http://www.dowservices.com/miss.html	100%	No terms
http://www.idsglobal.com	No terms	100%
http://www.idsglobal.com/Pack/Pkg_Nov_Member_info.htm	53%	No terms
http://www.homeopportunities.com	23%	100%
http://www.homeopportunities.com/security.htm	0%	No terms
http://www.clientconnectioninc.com	40%	No terms
http://www.clientconnectioninc.com/products.asp.html	30%	No terms
http://www.jinglebiz.com	53%	75%
http://www.jinglebiz.com/webpq2.htm	50%	No terms
http://www.camronnet.com	23%	100%
http://www.camronnet.com/net.html	No terms	No terms
http://www.elastix.com.au	71%	No terms
http://www.elastix.com.au/index16.htm	No terms	No terms
http://www.rayonyarn.com	61%	No terms
http://www.rayonyan.com/texapps/	50%	No terms
Average (does not include null results)	*47%*	*79%*

Table 2. Percentage of indexing terms rated most relevant by tool

The four sites for which the two indexing tools were both able to generate indexing terms were isolated to compare the ranked terms. Indexing terms generated by *Extractor* were considered the most pertinent (ranked first or second) for three of these sites. As before, these results are interesting and must be cautiously interpreted.

CONCLUSION

This pilot study has shown that an intra indexing tool seems to provide a richer set of keywords. *Extractor* generated considerably more terms, of which only half were considered to be the most relevant. Yet, an extra indexer seems to be able to derive more accurate terms, but not necessarily the best terms for a site. *Topic*'s terms, when they were generated, were always relevant but not always considered to be the most relevant. These results are very tentative and indicate that replication is essential before any definitive conclusions can be reached. Because of our site selection techniques (i.e., randomized at each decision point), we ended up with a set of sites that were unindexable by the tools and thus prevented a good comparative test.

REFERENCES

Chakrabarti, S., Dom, B., Gibson, D., Kleinberg J., Kumar, S.R., Raghavan, P., Rajagopalan S., & Tomkins A. 2000. Hypersearching the web. *Scientific American*. Retrieved September 26, 2000, from: http://www.scientificamerican.com/1999/0699issue/0699raghavan.html

Chakrabarti, S., van den Berg, M., & Dom, B. 1999. Focused crawling: a new approach to topic specific resource discovery. Retrieved September 18, 2000: http://www.cs.berkeley.edu/~soumen/doc/www1999f/html/

Keast, G., Toms, E.G., & Cherry, J. 2001. Measuring the reputation of web sites: a preliminary exploration. In the *First ACM/IEEE-CS Joint Conference on Digital Libraries, Roanoke, VA USA, June 24-28, 2001*. In press.

Kleinberg, J. M. 1998. Authoritative sources in a hyperlinked environment. *CACM*, 46 (5), 604-632. Retrieved September 20, 2000 from http://www.cs.cornell.edu/home/kleinber/auth.ps

Rafiei, D. & Mendelzon, A. O. (2000). What is this page known for? Computing Web page reputations. Retrieved September 6, 2000, from: ftp://ftp.db.toronto.edu/pub/papers/www9.ps.gz

Turney, P.D. 1997. Extraction of keyphrases from text: evaluation of four algorithms, National Research Council, Institute for Information Technology, Technical Report ERB-1051.

Turney, P.D. 2000. Learning algorithms for keyphrase extraction. *Information Retrieval*, 2 (4), 303-336.