8-2010

Forming Within-site Topical Information Space to Facilitate Online Free-Choice Learning

Ping Yan
University of Arizona, pyan@email.arizona.edu

Zhu Zhang
University of Arizona, zzhang@email.arizona.edu

Forming Within-site Topical Information Space to Facilitate Online Free-Choice Learning

Ping Yan
University of Arizona
pyan@email.arizona.edu

Zhu Zhang
University of Arizona
zzhang@email.arizona.edu

ABSTRACT

Locating specific and structured information in the World Wide Web (WWW) is becoming increasingly difficult, because of the rapid growth of the Web and the distributed nature of information. Although existing search engines do a good job in ranking web pages based on topical relevance, they provide limited assistance for free-choice learners to leverage the non-linear nature of information spaces for knowledge acquisition. We hypothesize that free-choice learners would benefit more from structured topical information spaces than a list of individual pages across multiple websites. We conceptualize a within-site topical information space as a sphere formed by linked pages centering on a web page. In this paper, we investigate techniques and heuristics to form the space. In particular, we propose a hybrid method that relies on not only content-based characteristics and user queries, but also a site's global structure. Experimental results show that consideration of website topology provides good improvement to page relevance estimation, indicating the clustering tendency of relevant pages.

Keywords

Topical crawling, focused crawling, topical information space, information retrieval, free-choice learners.

INTRODUCTION

The Web provides easy access to a vast amount of informational content to the average person; however, locating the information of user's interest has been one of the most challenging problems for internet usage. Popular search engines such as Google have tried to address this problem by ranking pages according to their relevance to a user's query and the page's popularity or authority. Search engines based on manually tagged directories such as Yahoo, instead rely on concept taxonomy to classify websites. In either case, a user relying on a general search engine typically types in their query terms, and then he gets in return tens or hundreds of web pages followed by a brief description of each page. The result set is a simple ranked list of URLs, each one of which leads to an information space up to the user's exploration. We believe that such an exploration process should be facilitated by automatically computing a bounded within-site topical information space consisting of multiple relevant pages, following the linkage structure of the website. The topical information space might be conceived as a sphere formed by linked pages centering on the user's requested information. We hypothesize that the information space can better serve a free-choice learner's searching objectives with a relatively complete and self-containing collection of clustered pages as opposed to a single page containing partial information.

However, a topical information space is naturally a vague notion. Many questions remain to be answered to select pages to form this information space: 1) Given a query, what pages in a site construct an information space? 2) How many pages should be included to form this space? 3) How to retrieve the pages? Which page to start with? And 4) How to measure the quality of the retrieved space? As a first attempt to answer these questions, several strategies for forming information spaces are proposed and evaluated.

The paper is organized as follows. We review related works in Section 2. In Section 3, the heuristics we use to develop the within-site topical information space are presented. Section 4 demonstrates the performance of the proposed methods. We conclude and present future work in Section 5.

RELATED WORK

The related literature covers three aspects in the broad area of information retrieval. They are page relevance measures, crawling strategies and results evaluation.

Page relevance measures A topical crawler makes relevance judgments on pages and decides on link expansion and visit priorities. Lexical relevance criteria and link-based relevance measure are proposed by previous researchers. Examples of lexical criteria include similarity between a page's vector and the seed documents [2, 3]; similarity between anchor text or neighborhood text, i.e., words surrounding a link and the seed documents [4]; and intuitively a linear combination of source page relevance, anchor text and neighborhood text relevance as relevance measure [8]. Link-based criteria consider in-degree, out-degree, such as page popularity measured by PageRank [15] and page authorities [10], in addition to text similarity of a page to the user query [5].

Crawling strategies People have proposed breadth-first [16], depth-first [6] or best-first search strategies [5, 8] to systematically explore graph structures such as the web. Variations exist by either letting a document inherit a discounted relevance score from its parent documents, proposed as "shark search"[8], or relying on a group of crawling agents work in parallel [12]. Existing studies suggest that best-first search strategy outperforms the other strategies [5, 8].

Evaluation Precision, recall, and F-measure are the common measures used to evaluate the performance of a web crawler. Precision basically measures the percentage of "good" pages retrieved over the progress of the crawl [5]. Recall measures the coverage of the retrieved collection on the target pages. F-measure is the harmonic mean of precision and recall. To compute these measures, a third-party evaluator is needed to evaluate page relevance, classifying pages as "relevant" and "non-relevant". The evaluator could be a human expert, a program evaluator, or a classifier.

INFORMATION SPACING ALGORITHM WITH QUERY HEURISTICS AND SITE CHARACTERISTICS

From a design science perspective, we design an artifact that consists of three components: a relevance evaluator, a crawling strategy, and a user interface. The user interface takes user search query as input. The search query summarizes the topics of user's interest. Meanwhile, it is optional for the user to also specify an entry URL of a website, where the user wants to form an information space for that website. The entry page can also be pre-determined by either a human expert, such as an experienced librarian, or suggested a trustable third party. A practical design could be that the user chooses from the top ranked pages suggested by a popular search engine. We start from the entry page, follow outgoing links, traverse the website, evaluate the relevance of pages encountered, and collect the most relevant pages to form a topical information space. Since best-first search has been shown by previous work to be the most effective crawling strategy, we focus our research on heuristic relevance functions.

Content-Based Relevance Evaluation

Using the anchor text pointing to a candidate page as its proxy is the most economic relevance measure, as there is no need to fetch the candidate document. In the traversal process, a candidate page is directly linked from the parent page, which is where the crawler is currently positioned, i.e., the last page that has been visited and retrieved as part of the information space. We denote the last page crawled in time $t-1$ as p_{t-1}, the candidate pages to be evaluated at time t is denoted as a set P_t. Anchor text associated to hyperlinks is a succinct representation of the content of a candidate page. Previous work suggests that anchor text indicates major content of a web page with high reliability [4, 7, 19]. It is also observed by Eiron et al that anchor text resembles user query in terms of length and term distribution, resulting in more coherent retrieval results [7]. At the same time, we extract visible content text from p_{t-1} removing any HTML tags, while keeping titles, headers, and text in tables, forms etc, for relevance evaluation.

p_t	Candidate page	Page to crawl at time t
p_{t-1}	Parent page	Page crawled at time t-1
T_q	Query	User input as desired topic
T_a	Anchor text	Succinct representation of candidate page's content
T_c	Content text	Page content text visible to web users
$\mathrm{sim}\left(T_a,\ T_q\right)$		Similarity with query string
$\mathrm{sim}\left(T_a,\ T_c\left(p_{t-1}\right)\right)$		Similarity with parent page's content text

Table 1: Notations of content-based heuristics related concepts

The similarity score between two pieces of text are computed as a cosine between two vectors in a m dimensional word space. Formally, similarity between items i and j, denoted by sim(i, j) is given by $sim(i, j) = \cos(\bar{i}, \bar{j}) = \frac{\bar{i} \cdot \bar{j}}{\|i\|^2 * \|j\|^2}$, where "$\cdot$" denotes the dot-product of the two vectors. It has been shown that the vector space model gives better results than simple string matching and regular-expression matching [17].

In addition, we test if a query similarity added to the similarity score between the candidate and the parent page improves information space forming. We hypothesize that the similarity between the candidate page and the query helps concentrate the traversal process on query relevance, while the one with parent page makes sure the next crawled page is relevant to the page just crawled.

Graph Partitioning Based Page Clustering As Crawling Heuristic

Besides the text similarity as crawling heuristics, we also examine the linkage structure of a website. In fact, concentration patterns can be observed from linkage graph of a website: most websites are organized as a collection of clustered pages [20]. Figure 1 depicts part of the linkage structure of the website of java.sun.com, showing that page tend to cluster from a global point of view. Web site linkage structure suggests important navigation layout, thus has potential in guiding crawlers. We propose to leverage the page clustering pattern observed from linkage graph of a website to guide our crawling process.

Graph is formed by extracting link structure from the web pages. We then use CLUTO [9] to perform graph partitioning, and thus result in page cluster membership. Link structure needs to be first converted to CLUTO graph format, where adjacency matrix specifies the connections between pages. The toolkit implements the min-cut graph partitioning algorithm for clustering vertexes on a graph. Figure illustrates a simplified two-way min-cut graph partitioning example.

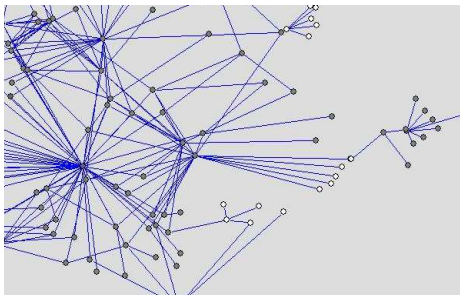

Figure 1: Pajek visualization of java.sun.com website's linkage structure

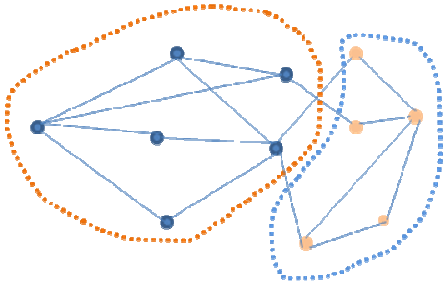

Figure 1: Example of 2-way min-cut graph partitioning

We leverage the cluster membership information generated from the graph-partitioning algorithm to further evaluate the page relevance. The page clustering heuristic is hypothesized to improve the crawler performance, by injecting the global landscape of page distribution into the crawling process. Each time, the crawled pages will be scanned to get the page distribution in each cluster. When there are n pages in the collection of crawled pages, suppose they belong to m clusters, and h_1, h_2, …, h_i, ..., and h_m (i is the cluster label) are number of pages belonging to each of the m clusters ($\sum_m (h_1, ..., h_m) = n$).

Weights w_1, w_2, and ...w_n are determined as: $w_i = e^{\frac{h_i}{n}}$, where w_i is a fraction between 1 and e. The page relevance is then adjusted by multiplying a weight, which corresponds to the cluster membership of that particular page.

To summarize, we have the following four models for information space formation (Table 2): 1) M1: in the crawling process, candidate page relevance is evaluated against the parent page; 2) M2: besides the similarity with its parent page, a page's similarity with the user query is considered also; 3) M3: clustering information is used to weigh page similarities; and 4) M4: incorporating relevance with both parent page and user query, as well as the page clustering information.

M1:w/o:q	$\text{sim}\left(T_a, T_c\left(p_{t-1}\right)\right)$

M2:w/:q	$\text{sim}\left(T_a,\ T_c\left(p_{t-1}\right)\right)+\ \text{sim}\left(T_a,\ T_q\right)$
M3:w/:cl	$\text{sim}\left(T_a,\ T_c\left(p_{t-1}\right)\right)*w_i$
M4:w/:cl:w/:q	$\left(\text{sim}\left(T_a,\ T_c\left(p_{t-1}\right)\right)+\ \text{sim}\left(T_a,\ T_q\right)\right)*w_i$

Table 2: Formulation of four information space forming models

EXPERIMENTS AND RESULTS

Test queries can be generated in several ways: traverse a topic directory structure such as Yahoo or Open Directory Project (ODP), and use either the leaf node concepts as topics [13] or build them from sub-trees of a given maximum depth [4]. We randomly select our query strings from the hierarchical concept directories of ODP. Noticing that query characteristics might affect crawler's performance [18], we picked five diverse tropics covering science, technologies, sports, economics and art. These queries are: "object-oriented programming", "art British literature", "sports fencing", "financial services mortgage", and "science physics astronomy". For each query, we developed information spaces for 10 top-ranking URLs returned by Google. Each website derived from the URLs originally contains pages numbering from a dozen to a few hundreds.

Lemur is a state-of-the-art information retrieval system in the academic area. We used its text search engine-Indri [1, 11], developed at UMass, to retrieve relevant pages within each site as the gold standard collection. After Lemur indexing application run through the document collection, we submit the user query to Indri retrieval engine, Indri uses the language modeling approach to assign a relevance score to each document, where the score is the log of a probability value that is between 0 and 1 [14]. The Indri retrieval engine replies on an inference network that is made up of document nodes (from the indexed collection), smoothing parameter nodes, representation concept nodes, language model nodes, belief nodes, and information (or combination of information) nodes. It retrieves documents with non-zero relevance scores (not exceeding a maximum of 1000 pages). And we view the retrieved document collection as a gold standard collection of relevant pages.

Precision, recall and F-measure are used as evaluation metrics. Table 3 shows all three measures varying by different information space sizes (the number of pages in the space). Each row corresponds to a space size, ranging from 10 to 50. As the size of information space increases, recall increases across all four models, while precision drops slightly. Based on the T-Test resutlts shown in Table 3: Recall, precision and F-measure of the four information space forming models

, M4 is significantly better than M2 at the .10 level yet only marginally better than M3, which indicates that the query information is not as useful as the global landscape manifested in page clusters. This is possibly due to the high overlap between the query and the seed page, since the latter is returned by Google in response to the query and therefore typically a good approximation of the query.

	M1:w/o:q			**M2:w/:q**			**M3:w/:cl**			**M4:w/:q:w/:cl**		
	Rec	**Pr**	**F**	**Rec**	**Pr**	**F**	**Rec**	**Pr**	**F**	**Rec**	**Pr**	**F**
10	11.6%	22.2%	7.3%	14.2%	41.8%	11.5%	20.0%	46.5%	19.6%	21.4%	50.9%	19.8%
20	14.4%	16.3%	7.2%	18.7%	35.9%	13.0%	23.6%	39.3%	20.5%	25.2%	42.8%	20.8%
30	15.3%	15.0%	7.6%	20.4%	32.2%	14.0%	26.0%	36.6%	21.3%	27.2%	39.2%	21.9%
40	15.9%	14.6%	7.9%	22.5%	29.5%	14.5%	26.5%	34.5%	21.5%	28.0%	37.1%	22.3%
50	16.2%	14.0%	8.1%	23.9%	27.8%	14.8%	27.2%	33.6%	21.9%	28.6%	35.2%	22.5%

Table 3: Recall, precision and F-measure of the four information space forming models

Size	**M2>M1**	**M3>M1**	**M4>M1**	**M3>M2**	**M4>M2**	**M4>M3**
10	0.0%	0.1%	0.1%	2.7%	2.6%	22.3%
20	0.1%	0.1%	6.3%	7.0%	6.3%	4.9%
30	0.0%	0.1%	6.5%	7.9%	6.5%	10.6%
40	0.0%	0.1%	0.1%	10.2%	7.4%	9.5%
50	0.0%	0.1%	0.1%	9.6%	7.6%	11.0%

Table 4: Model comparisons: p values for T-tests

CONCLUSION AND FUTURE WORK

Free-choice learners seeking information on the Internet need more structured and self-contained collection of web pages for learning purpose, while existing search engines cannot meet this requirement by returning the user a list of separate web pages scattered on the Web. We proposed a framework to develop a within-site topical information space to assist free-choice learners in exploring a website for more effective knowledge acquisition. The proposed approach retrieves web pages deemed relevant to the user query, not only replying on the content-based relevance, also leveraging the site's topological characteristics.

Experimental results show that website topology, i.e. linkage structure, provides good improvement to page relevance estimation, due to the clustering tendency of relevant pages. Built upon the preliminary study presented here, in the future, we will investigate other probabilistic weighing algorithms to guide the crawling process more intelligently. In addition, instead of using a third-party document retrieval engine as the gold standard, we consider designing human subject experiments to investigate the effects of a within-site topical information space in online free-choice learning processes

REFERENCES

1. *Indri Retrieval Model Overview*. 2005: http://ciir.cs.umass.edu/~metzler/indriretmodel.html.
2. Amento, B., L. Terveen, and W. Hill. *Does "authority" mean quality? Predicting expert quality ratings of web documents*. in *the 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2000.
3. Bharat, K. and M. Henzinger. *Improved algorithms for topic distillation in hyperlinked environments*. in *the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1998.
4. Chakrabarti, S., M.v.d. Berg, and B. Dom. *Focused crawling: A new approach to topic-specific web resource discovery*. in *the 8th International World Wide Web Conference*. 1999.
5. Cho, J., H. Garcia-Molina, and L. Page. *Efficient crawling through URL ordering*. in *the Seventh International World Wide Web Conference*. 1998. Brisbane, Australia.
6. De Bra, P. and R. Post. *Information retrieval in the World Wide Web: Making client-based searching feasible*. in *1st International World Wide Web Conference*. 1994. Geneva.
7. Eiron, N. and K.S. McCurley. *Analysis of anchor text for web search*. in *the 26th ACM SIGIR Conference*. 2003.
8. Hersovici, M., et al. *The Shark-Search Algorithm - An Application: Tailored web Site Mapping*. in *the Seventh International World Wide web Conference*. 1998.
9. Karypis, G., *CLUTO: Software package for clustering high dimensional data*. 2002: www.cs.umn.edu/~karypis/cluto
10. Kleinberg, J., *Authoritative sources in a hyperlinked environment.* Journal of the ACM 1999. **46:5**(5): p. 604-632.
11. Lavrenko, V. and W.B. Croft. *Relevance-Based Language Models*. in *the 24th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (SIGIR '01)*. 2001.
12. Menczer, F. and R. Belew, *Adaptive retrieval agents: Internalizing local context and scaling up to the web.* Machine Learning, 2000. **39**: p. 203-242.
13. Menczer, F., et al. *Evaluating topic-driven Web crawlers*. in *the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2001. New York, NY: ACM Press.
14. Metzler, D. and W.B. Croft, *Combining the Language Model and Inference Network Approaches to Retrieval.* Information Processing and Management Special Issue on Bayesian Networks and Information Retrieval, 2004. **40**(5): p. 735-750.
15. Page, L. and S. Brin. *The anatomy of a large-scale hypertext web search engine*. in *the seventh International World Wide Web Conference*. 1998.
16. Pinkerton, B. *Finding what people want: Experiences with the WebCrawler*. in *the First World Wide Web Conference*. 1994. Geneva, Switzerland.
17. Salton, G., *Automatic Text Processing, the Transformation, Analysis and Retrieval of Information by Computer*. 1989, Reading, MA: Addison-Wesley.
18. Srinivasan, P., F. Menczer, and G. Pant, *A General Evaluation Framework for Topical Crawlers.* Information Retrieval, 2005. **8**(3): p. 417-447.
19. Tombros, A. and Z. Ali, *Factors Affecting Web Page Similarity*, in *ECIR 2005, LNCS 3408*, D.E. Losada and J.M. Fern´andez-Luna, Editors. 2005, Springer-Verlag Berlin Heidelberg. p. 487–501.
20. Zeiliger, R. *Supporting Constructive Navigation of Web Space*. in *the Workshop on Personalized and Solid Navigation in Information Space*. 1998.