

On the Topic Discovery Using Query Logs and Hyperlink

Chih-Ming Tseng

Department of Information
Management
National Taiwan University of
Science and Technology
Taipei, Taiwan

Yun-Fei Wei

Department of Information
Management
National Taiwan University of
Science and Technology
Taipei, Taiwan

Chiun-Chieh Hsu

Department of Information
Management
National Taiwan University of
Science and Technology
Taipei, Taiwan

Abstract. With the rapid growth of the World Wide Web, the amount of information in the Web has spawned on an unpredictable scale. Recently, most researchers attempt to use conventional information retrieval techniques to classify the search results. But not like the traditional document, the web page has distinct characteristics of its own. Therefore some researchers have begun to exploit the hyperlinks between Web pages. In this paper, we propose a topic discovery algorithm, which combines the query log and the hyperlink analysis. We use the query log to find the representative Web pages with respect to users' endorsements, and combine a link-based clustering algorithm to cluster the similar topics. Each Web page is ranked according to search engine users' endorsements and web creators' endorsements. The experimental results show that our method performs better than the pure hyperlink analysis algorithm (ATD) in terms of topics discrimination and topic quality.

Keywords: Query Log, Topic Discovery, Hyperlink.

1 Introduction

Traditional Web page search is based on user's query terms and Web search engines, such as AltaVista [1] and Google [3]. The problem with search engines is that they often return too many results, and document ranking may not meet the user's expectations [2]. In addition, many search engines only return a set of individual documents. Internet documents returned from a search engine may contain several topics about the input query. For example, the results of query "briefcase" may contain topics such as "Yahoo! briefcase" and "products, bag, briefcase", etc. It is obvious that the documents from different topics represent different interests. For web document retrieval, it helps users to assimilate the results by partitioning the documents into topics and annotating each topic. Furthermore, it is

sometimes important to rank the documents according to their importance to a specific topic.

Topic discovery is an emerging technology that can be applied to enhance search engine results. Wu [5] propose an algorithm ATD that partitions the web pages in search results into clusters, and rank each topic according to its authority. In this paper, we propose a topic discovery algorithm that combines the query log and web page hyperlink analysis, and develop a system called TDQL (Topic Discovery algorithm for Query Log and web page hyperlink analysis). We use the query log to find the representative Web pages with respect to users' endorsements, and combine a link-based clustering algorithm to cluster the similar topics. Each web page is ranked according to search engine users' endorsements and web creators' endorsements. The experimental results show that our method performs better than the pure hyperlink analysis algorithm ATD in terms of topics discrimination and topics quality.

The rest of this paper is organized as follows. Section 2, gives a short review of ATD. Section 3 provides detail about TDQL algorithm. Section 4 shows the results concluded from experiments. Finally, section 5 concludes the paper.

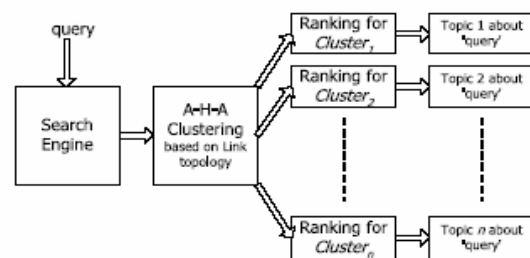


Fig. 1. Automatic Topic Discovery (ATD)

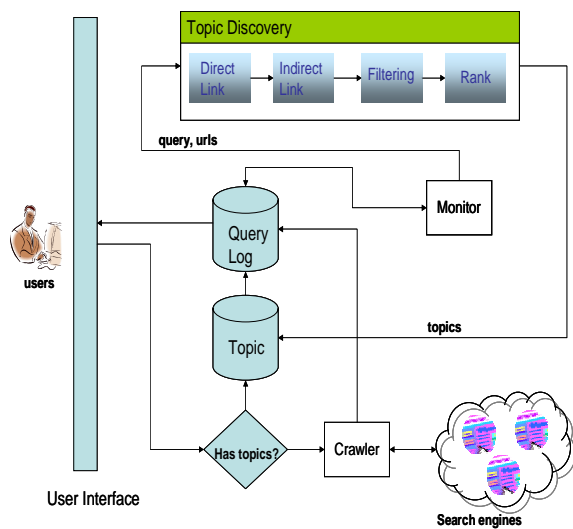


Fig. 2. TDQL system architecture



Fig. 3. TDQL User Interface

2 Review of ATD

The aim of the ATD algorithm is to identify and isolate each strongly inter-connected cluster as topic in the web vicinity graph, and then select top-ranked web pages within each cluster to be its representing concept. The complete workflow of automatic topic discovery is illustrated in Fig. 1. First a broad-topic query is sent to a content-based search engine. The content-based search engine searches its index against the query and returns a list of matched documents. Based on the search results, page contents are examined to obtain link information. Then, an expanded focused web vicinity graph, composed of relevant web pages and hyperlinks, is built. By following the underlying hyperlink topology and relationship between authority and hub, the vicinity graph is partitioned into several inter-connected clusters relevant to the broad-topic query. These clusters are ranked by an eigenvector computing algorithm.

3 TDQL system

The TDQL system is shown in Fig. 2. The system includes following components.

3.1 User Interface

User can input a broad-topic query through the system interface as shown in Fig.3. When user submit query string in the user interface. The TDQL system retrieval the documents relevant with query string in the topic database, and combine the returned by web crawler results. The web crawler collect web documents relevant with query string, which were returned by result of search engine, e.g. AltaVista and Google.

3.2 Database & Monitor

It includes topic database and query log database. Topic database was stored different clustering web pages, which were returned by results of search engine. query log database was stored log of query string and log of user's clicked web pages, called session count. The monitor monitors change of query log. When support of session count is greater than predefined threshold value, the monitor automatically triggers Topic Discovery mechanism to produce topic clustering web pages association with query string.

3.3 Topic Discovery mechanism

The aim of the Topic Discovery mechanism is to identify and isolate each strongly inter-connected cluster as topic, and then select top-ranked web pages within each cluster. The Topic Discovery includes two parts: clustering and pages ranking.

3.3.1 Clustering.

Kleinberg [4] proposed that every web page has two properties: authority and hub. A web page's authority is promoted if linked by many pages. The hub attribute indicates the quantity of the hyperlinks which link to high quality authority pages contained in the web page content. In topic discovery, a partitioned cluster that contains a topic is composed of good authority and hub pages. To exploit the above concept to discover topic we first use that the support of clicked web pages are association with query string is greater than pre-defined threshold value, then they are representative web pages for query string to construct a base set. Second, to use HIT method that expands the base set according to link information.

Table 1. Topic discovery result of query “Jaguar”, TDQL and ATD method

Topic	The result of TDQL method	The result of ATD method
Racing	http://www.jaguarracing.com	http://forum.planet-f1.com/index.php
	http://www.usgpindy.com	http://www.bettingzone.co.uk
	http://www.f1total.com	http://www.sportal.com
Club	http://www.jaguar-association.de	http://www.jcglv.org/links.htm
	http://www.jaguar.org.au	http://forums.jag-lovers.org
Animal	http://dSPACE.dial.pipex.com/agarman/bco/jaguar.htm	http://www.animail.com/posters_jaguar.html
	http://www.ucs.louisiana.edu/~csn1234/1nationlist.htm	http://www.art.com/asp/sp.asp?pd=10034105&rfid=642275
Car	http://www.audi-autoparts.com	http://www.audi-autoparts.com
	http://www.lexus-autoparts.com	http://www.lexus-autoparts.com
	http://www.buick-autoparts.com	http://www.buick-autoparts.com
Civilization	http://www.jaguar-sun.com	
	http://www.civilization.ca/civil/maya/mmj01eng.html	
Band	http://www.jaguar-online.com	
	http://www.unbrokenmetal.de/umm/links.htm	

Pages that are linked by any pages in the base set or web pages which link to any pages in the base set are added into the base set. Third, we cluster different topics using direct link relationship and indirect link relationship. In the indirect link relationship that we select a web page with maximum in-degree as centric page in every topic, and use HIT algorithm to calculate authority and hub that any pages were linked by centric page or centric page which link to any pages, then select top 20 authorities and 15 hubs web pages are added into topic. A cluster is discarded if the number of web pages in that cluster is less than a pre-defined threshold value.

3.3.2 Pages ranking.

The rank of web pages are combined session count and web page in-degree, it's calculate expression is shown in equation (1).

$$A_{rank} = \alpha |A_{Session}| + (1 - \alpha) |A_{in\ degree}| \quad (1)$$

$|A_{session}|$: session count

$|A_{indegree}|$: the count of web page A with in-degree

α : tunable parameter, $0 \leq \alpha \leq 1$.

4 Experiments

In this experiment, we use an Intel Pentium 3.0GHz PC. We use two web collections. First, we use TDQL to record user search log and the log of the user clicked web pages association with query string. When

user input query string in TDQL, and the system submit query string to the search engine, then the search engine return the results of the query string and through TDQL represents to user. Second, we collect 1,099,976 query logs of the National Taiwan University of Science and Technology's proxy server from May 2004 to September 2004. First, we evaluate topic discovery result of query “Jaguar” for TDQL and ATD. In Table 1, the result shows the topic number of the ATD is less than TDQL. Next, we evaluate topic relevance for TDQL and ATD.

We recruited three volunteers to rate the topics. The rating criteria includes following choices: 1 stands for high relevance, between 0 and 1 stands for relevance, and 0 stands for non-relevance. Experiment results are shown in Table 2. The experimental results show that our method performs better than the pure hyperlink analysis algorithm (ATD) in terms of topics discrimination and topics quality.

Table 2. Topic quality, TDQL method and ATD method

Topic ID		1	2	3	4	5	6	7	8	Avg
Jaguar	TDQL	1	1	1	0.7	0.8	0.9	1	*	0.91
	ATD	0.5	0.5	0.9	0.67	0.3	*	*	*	0.57
Tiger	TDQL	1	1	0.7	1	1	0.6	1	1	0.91
	ATD	0.33	0.8	0.14	0.3	0.7	0.9	1	0.5	0.58
Apple	TDQL	1	0.9	0.9	0.7	*	*	*	*	0.88
	ATD	0.8	1	0.9	0.7	*	*	*	*	0.85
Matrix	TDQL	1	0.7	0.9	0.7	0.8	0.6	0.6	0.7	0.63
	ATD	0.6	0.14	0.38	*	*	*	*	*	0.37

5 Conclusion

In this paper, we propose a topic discovery algorithm that combines the query log and web page hyperlink analysis, and develop a system called TDQL (Topic Discovery algorithm for Query Log and web page hyperlink analysis). TDQL is off-line computing algorithm that doesn't affect total efficiency and all data were stored in the database for future searching. When search logs are more and more, the TDQL can identify more topics and related of intra-document is high relevance in every topic. We also propose new web page ranking method; it combines users' view and web page designers' view. The experimental results show that our method performs better than the pure hyperlink analysis algorithm (ATD) in terms of topics discrimination and topics quality.

6 References

- [1] AltaVista, <http://www.altavista.com/>
- [2] Chaffee, J., & Gauch, S. "Personal ontologies for web navigation" In Proceedings of the 9th international conference on information and knowledge management, McLean, VA (2000), 227-234.
- [3] Google, <http://www.google.com/>
- [4] Kleinberg, J. M, "Authoritative Sources in a Hyperlinked Environment," In Proceedings of the 9th annual ACM-SIAM symposium on discrete algorithms, San Francisco, CA, (1998) 668-677.
- [5] Kuo-Jui Wu, Meng-Chang Chen, Yeali Sun. "Automatic topics discovery from hyperlinked documents," Information Processing and Management 40(2004), 239-255.