Available online at www.sciencedirect.com
ScienceDirect
Information Processing and Management 43 (2007) 445–460
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Topic distillation via sub-site retrieval ☆

Tao Qin [a,*,1], Tie-Yan Liu [b,2], Xu-Dong Zhang [a], Guang Feng [a], De-Sheng Wang [a], Wei-Ying Ma [b,2]

[a] *MSP Laboratory, Department of Electronic Engineering, Tsinghua University, Beijing 100084, PR China*
[b] *Microsoft Research Asia, No. 49, Zhichun Road, Haidian District, Beijing 100080, PR China*

Received 15 June 2006; accepted 25 July 2006
Available online 11 October 2006

Abstract

Topic distillation is one of the main information needs when users search the Web. Previous approaches for topic distillation treat single page as the basic searching unit, which has not fully utilized the structure information of the Web. In this paper, we propose a novel concept for topic distillation, named sub-site retrieval, in which the basic searching unit is sub-site instead of single page. A sub-site is the subset of a website, consisting of a structural collection of pages. The key of sub-site retrieval includes (1) extracting effective features for the representation of a sub-site using both the content and structure information, (2) delivering the sub-site-based retrieval results with a friendly and informative user interface. For the first point, we propose Punished Integration algorithm, which is based on the modeling of the growth of websites. For the second point, we design a user interface to better illustrate the search results of sub-site retrieval. Testing on the topic distillation task of TREC 2003 and 2004, sub-site retrieval leads to significant improvement of retrieval performance over the previous methods based on single pages. Furthermore, time complexity analysis shows that sub-site retrieval can be integrated into the index component of search engines.

Keywords: Sub-site retrieval; Topic distillation; Website expansion model; Punished-integration algorithm; User interface

1. Introduction

While the World Wide Web grows exponentially along with time, the amount of information that users can digest remains roughly constant. With such background, search engines came out as the widely recognized tools in helping users retrieve information from the Web. According to some previous surveys, among all

☆ This work was performed when the first and the fourth authors were interns at Microsoft Research Asia.

* Corresponding author. Tel.: +86 10 6278 9944.
E-mail addresses: qinshitao99@mails.thu.edu.cn (T. Qin), tyliu@microsoft.com (T.-Y. Liu), zhangxd@tsinghua.edu.cn (X.-D. Zhang), fengg03@mails.thu.edu.cn (G. Feng), wangdsh_ee@tsinghua.edu.cn (D.-S. Wang), wyma@microsoft.com (W.-Y. Ma).

[1] The work was supported in part by the Joint Key Lab on Media and Network Technology set up by Microsoft and Chinese Ministry of Education in Tsinghua University.

[2] Tel.: +86 10 6261 7711.

0306-4573/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2006.07.004

the information needs of Web users, topic distillation is one of the most common and important forms. Here, topic distillation refers to the search scenario of distilling a small number of high-quality entry pages that are best representatives of a given broad topic. It was shown (Broder, 2002) that about 39% Web search queries belong to topic distillation, while query log analysis has indicated an even higher proportion of 48%. Because of the popularity of topic distillation, the TREC conference has included it in the Web track since the year of 1999.

To the best of our knowledge, in most existing approaches for topic distillation, the single page was treated as the basic searching unit. Moreover, it has been a common practice in the literature of topic distillation to use some statistics of the query terms in the page content (such as TF-IDF (Baeza-Yates & Ribeiro-Neto, 1999)) to compute a relevance score (using information retrieval models such as BM25 (Robertson, 1997)) and use hyperlinks to get an importance score (such as PageRank (Page, Brin, Motwani, & Winograd, 1998)), then combine them to rank the retrieved web pages. However, this strategy has not yet been successful enough for topic distillation. According to the report of TREC 2003 (Craswell & Hawking, 2003), the best result produced by this strategy only has marginal retrieval accuracy: precision at 10 (P@10) of 12.8% and mean average prevision (MAP) of 15.43%. In other words, there is still a long way to go if one wants to get satisfactory search results for topic-distillation queries.

In order to improve the search performance for topic distillation, we investigated the ground truth provided by the TREC committee. Our observation is that the labeled positive answer for topic distillation is often the entry point of a collection of pages devoted to the query topic. That is, for topic distillation, whether a page is an appropriate answer to the query is not only determined by the page itself, but also by all the other pages for which it serves as the entry point. In this regard, topic distillation is very different from traditional information retrieval. Actually, its objective is not to find a single page but a group of pages, although only the entry point of the page group is labeled as positive answer. In other words, in order to improve the retrieval performance of topic distillation, we need to take the structural relationship between web pages (i.e., the website structure) into consideration.

As we know, the organizational structure of a website is usually represented by a tree (or sitemap), and the parent–child relationship between any two pages can be derived from sitemap construction (Qin, Liu, Zhang, Chen, & Ma, 2005). In this way, one can get all the descendent pages for a labeled entry page. According to our analysis, in many cases, some descendant pages are even more relevant than the labeled entry page in terms of TF-IDF (Baeza-Yates & Ribeiro-Neto, 1999). The reason is that the labeled entry page may just be a navigation center with only a few words while concrete contents are placed in its descendant pages. Take Table 1 for example, there are totally seven pages coming from the same site (http://cio.doe.gov/) in the top 1000 retrieval results produced by the BM25 model (Robertson, 1997) for the query "wireless communication" of the topic distillation task in TREC 2003 Web track. If we base our ranking on the content relevance of single pages, it will be very difficult to rank the labeled positive page "cio.doe.gov/wireless/" to the top one position because this page appears not as relevant to the query as its descendant pages "cio.doe.gov/wireless/3g/3g_index.htm", "cio.doe.gov/wireless/background.htm", and "cio.doe.gov/wireless/wwg/wwg_index.htm".

To solve the aforementioned problem, we propose to change the basic searching units from single pages to sub-sites. Here, sub-site refers to the sub-tree of a website, which contains an entry page and all its descendent

Table 1
Retrieval results for the query "wireless communication"

Rank	Document ID	Relevance Score	URL
70	G35-97-1056561	9.858	cio.doe.gov/wireless/3g/3g_index.htm
470	G07-38-3990160	9.508	cio.doe.gov/spectrum/groups.htm
477	G35-75-1119753	9.481	cio.doe.gov/spectrum/philo.htm
518	G36-35-1278614	9.320	cio.doe.gov/wireless/background.htm
571	G07-10-2999356	9.093	cio.doe.gov/spectrum/background.htm
648	G35-01-1537522	8.817	cio.doe.gov/wireless/wwg/wwg_index.htm
649	**G07-78-3824915**	**8.815**	**cio.doe.gov/wireless/**
...	...	...	...