University of Ottawa at TREC 2010 Web Track

Ranking Web Documents Using Meta-Data

Falah Al-akashi and Diana Inkpen
University of Ottawa
Canada
falak081@uottawa.ca and diana@site.uottawa.ca

ABSTRACT

This paper describes the details of our participation in the Web track of the TREC 2010. Our approach collects information from the meta data and from some html tags of the web pages. Meta data is often used by the authors to describe the main topic or purpose of the webpage. Usually, the index words are collected from the visible data, but our method is preferable, when the webpage contains only multimedia data, such as images or animations; when the webpage contains only URL links or very little text data (e.g., home pages); when the webpage contains several different topics; when documents are repetitive; when we want to reduce the size of the inverted index; and when the weighting of a webpage depends on specific topics, not on word distribution. We indexed only selected sections of each webpage: "URL", "Title", "Meta", "Header", and "Alt" fields. The "Header" field is the only visible text field that we used.

1. INTRODUCTION

Meta-data ranking is a new method used in web document retrieval. We present a system which adds a new leverage to a standard retrieval technique by using only meta-data in order to produce a fast web search engine.

We built our system on the documents collection named Category B, which is a sub-part of ClueWeb10 collection. This was our first participation in the TREC Web Track. Our system focused on the ad-hoc task, while trying to also ensure diversity of the retrieved documents.

The rest of this paper is organized as follows. In Section 2, we describe the collection processing for both indexing and ranking. In Section 3, we discuss our

experimental results. Related work is discussed in Section 4. Finally, we conclude our method in Section 5.

2. COLLECTION PROCESSING

2.1 Indexing and querying ClueWeb10

We present a new method for indexing the ClueWeb10 collection, with focus on meta-data. We used simple stemming and stop-word removal at the query time. However, we integrated the stemming step between queries and the indexed data by applying a simple strategy for adding singular to the words if there were any plurals. In addition, we used a stop-word list with 430 words[1] to exclude stop-words from queries. First, we used our system with the 50 provided training queries. We extracted meta data from web pages in order to build xml files. Finally, we integrated the ranking step between our approach for indexing and ranking meta data. We used the Apache Lucene.net library for enhancing our ranking and query processing[2]. In order to have more flexibility in developing our approach for future purposes and also for better performance, we built three index files:

- Index containing meta data, in addition to URLs and TREC-ID of all pages from ClueWeb10 (Category B) collection, except those contained in the bundles named “enwpXX",
- Index containing URLs, titles, TREC-IDs of all pages from the bundles named “enwpXX" (the Wikipedia pages).
- Index at query time containing singular URL from each domain (site), including the pages from sub-domains.

The reason why we used these three indexes is that the first index is for retrieving only one Wikipedia page for each query; whereas the second index is for retrieving as many documents as possible with high ranks. The third index, we aimed to use it for the diversity task, in which URLs are used to filter the results such by domain names. If there is more than one document from the same site or domain, we move it lower on the ranking list.

[1] http://www.site.uottawa.ca/~diana/csi5180/StopWords

2 http://lucene.apache.org/java/docs/index.html

2.2 WIKIPEDIA COLLECTION RANKING

Wikipedia is an online encyclopaedia that recently becomes one of the largest repositories of knowledge, with millions of articles available for a number of languages. English Wikipedia is a part of the ClueWeb10 (Category A & B) and represents a spam-free collection of web-pages with dedicated descriptions of around 3,500,000 concepts or articles. While it is not possible to find any article a user may ask for among of those described in Wikipedia, there is still a very high chance that most potentially popular queries can be answered by ranking articles described in this repository. The Wikipedia part of the ClueWeb10 is a collection of HTML pages; therefore it needs simple basic filtering in order to serve as a web repository. Since each article in Wikipedia covers one topic, we did not index the entire content of the documents. Our system indexed only the title sections. We also ignored non-article pages such as lists, files, disambiguation pages, category pages, etc. at indexing time. We used only the page retrieved with the highest rank for each query, in order to avoid the problem of duplicating documents in the answers, because sometimes Wikipedia uses different concepts to name the same article, for example (*Airline* and *Aircraft*) or (*Malware* and *Computer Virus*).

2.3 WEB COLLECTION RANKING

Another possible source for finding relevant results for a particular query is ranking the main part of the ClueWeb10 web collection (excluding the Wikipedia articles). Basically, web pages are composed of different attributes. Each attribute is used for a specific purpose; for instance, some of them were used to archive the data or to help our search engine to distinguish to which category a page belongs to. Other attributes give simple descriptions for each web page. Our system assigned each attribute a specific weight, from the highest weight for URLs, down to the lowest weight for the Alt fields. In total, we used five attributes for indexing each web page: URL, Title, header (H1), keywords-description, and Alt. Our system used the following formula in order to rank a document regarding to a query:

$$SC(q,d)=\sum_{i=1}^{n} z_i \left(\frac{\sum_{t=1}^{y} pt}{gz_i} \right) + c^* \log\left(\sum_{t=1}^{x}\sum_{i=1}^{y} qt\ zi + 5\right); \quad 0 \le SC(q,d) \le 1$$

where n = number of attributes, except Alt.

x = number of attributes in the document.

y = number of words in the query.

qt = parametric value: if the term *t* appears in the attribute Alt.

zi = 1 , if the term appears in an attribute; 0 otherwise.

pt = *parametric* index for the position of the term *t* in the query.

gz_i = weight for attribute *i* if the query term is available in that HTML tag

c = complementary factor (0.22)

For efficiency, we assigned 1 in the case of the existence of a term in any attribute, and we ignored using 0 values in the computation, in case the term was absent from the particular attribute.

3 EXPERIMENTS AND RESULTS

The documents from the Category B part of the collection are provided as HTML raw pages; therefore we pre-processed the dataset, page by page, as follows. We extracted the content of the fields WARC-TREC-ID, WARC-Target-URL, and some important attributes from HTML such as title, header H1, meta keywords, meta description, and alt tags, in order to generate XML files. These HTML tags Were used in our module for indexing the documents. Then we used the Lucence library to enhance our document ranking. Stop-words were removed at query time.

We submitted only one run, DFalah2010, as a first attempt. The results of our submission run are given in table 1 as assessed by the track organizers; we present the best result (for the query that obtained the best result) and the mean score of the 36 test queries, for MAP (Mean Average Precision), P@k (precision at top k result), as well as the other measures reported by the organizers.

We also present our results in comparison to the average over all the 88 runs submitted by all the participants, for the test queries (36). Our experimental results in comparison to the average of the TREC 2010 Web Track are shown in table 2. Table 3 compares the best results over the test queries.

DFalah2010	MAP	P@5	P@10	P@20
Best	0.0465	0.6000	0.8000	0.5000
Mean	0.0110	0.1556	0.1333	0.1194

DFalah2010	MAP-IA	αDCG@5	αDCG@10	αDCG@20
Best	0.044597	0.593209	0.585290	0.668684
Mean	0.008163	0.158280	0.178036	0.213085

DFalah2010	αnDCG@5	αnDCG@10	αnDCG@20
Best	0.751802	0.684204	0.672716
Mean	0.190764	0.205070	0.243339

DFalah2010	P-IA@5	P-IA@10	P-IA@20
Best	0.333333	0.225000	0.150000
Mean	0.071944	0.051296	0.044931

Table 1: Results of our web track submissions

Run-ID	**αnDCG@10**	**IA-P@10**
DFalah	0.205	0.051
Average over 88 runs	0.213	0.083

Table 2: Average mean results (88 runs/36 test queries) and our submission

Run-ID	**αnDCG@10**	**IA-P@10**
DFalah	0.684	0.225
The best of the 88 runs	0.586	0.302

Table 3: The best results (36 test queries) and our submission

4. RELATED WORK

The system we presented and the underlying data discovery process uses meta data in order to select data from web pages. Web pages have specific characteristics, compared to standard text documents, challenging methods that are not biased towards these characteristics (such as anchor text, link graph, meta-data, etc.). There is no agreement in the literatures on the best scheme for making use of anchor text. Early attempts exploited anchor texts to index the web pages where the links point, according tp Brin and Page (1998) [13], McBryan (1994) [14], and Fujita (2001) [15]. Singhal and Kaszkiel (2001) [16] could not observe any reliable improvement in retrieval effectiveness when using the anchor texts.

Another characteristic of web documents to be considered is the document structure. The web documents have the title section and several levels of header sections (H, standing for headlines). From early web IR research, systems attempted to utilize this structure, but their importance was not clear, because using the structure did not result in enhancement in retrieval effectiveness in the topic relevance task (Amati & Carpineto, 2002; Savoy & Picard, 2001) [17]. However, more recent systems tend to use the title as a major document representative, especially in the named homepage finding task (Craswell & Hawking, 2003) [18].

URL is another source of information for web retrieval. It was shown that its information could be valuable in the home page finding task (Fujita, 2002) [19]. However, it was not confirmed that it is helpful in the named page finding task, according to other researchers (Craswell & Hawking, 2003) [18]. URL structures also include features that can help in web topic classification (Baykan, Henzinger, Marian, and Weber, 2009) [5].

Some studies proved that the index became fairly expensive to maintain and update frequently; therefore, meta data (title, anchor text, meta fields), and the texts in paragraph tags (between and) for ranking pages is preferable for indexing (Hodog Li, 2003) [20].

5. CONCLUSIONS

This paper describes the method and the experiments of our participation in ad-hoc and the diversity tasks in Web Track at TREC 2010. For these tasks, we used a new approach for indexing the important attributes on the documents, mostly contain non-visible data (meta data), except for web page titles. Additionally, we used a new model with the different weights for ranking the significant attributes in the documents that are completely different from the models described in the

previous section. As first attempt, our system obtained acceptable results. Our system did not use spam filtering; therefore the fact that we retrieved spam documents in each result degraded our total results, in some cases.

In future work, we will refine our model to include spam filtering and we will design a new approach to improve the efficiency of the web page ranking and the speed of the retrieval.

6. ACKNOWLEDGEMENTS

Thanks to the TREC 2010 Web track staff for shipping the dataset fast; also, thanks to the NIST assessors for evaluating the submission results.

RERFERENCES

[1] Arul Prakash Asirvatham, Kranthi Kumar Ravi, "Web page categorization based on document structure", Centre for Visual Information Technology, Gachibowli, INDIA.

[2] David A. Grossman, Ophir Frieder, "Information Retrieval Algorithms and Heuristics", Illinois Institute of Information Technology, Chicago, 2004.

[3] Christopher D Manning, Prabhakar Raghavan, Hinrich Schutze, "Introduction to Information Retrieval", Cambridge University Press, 2000.

[4] Sergey Brin and Lawrence Page, "The anatomy of a large-scale hyper textual web search Engine", In Proceedings of the Seventh International World Wide Web Conference, Brisbane, Australia, April 1998.

[5] Eda Baykan, Monika Henzinger, Ludmila Maria, Lngmar Weber, "Purely URL based Topic Classification", Lausanne Switzerland, 2009.

[6] Marko Grobelink and Dunja Mladenic, 'Simple Classification into Larger Topic Ontology of Web Documents", journal of Computing and Information Technology, Jozef Stefan Institute, Slovenia 2005.

[7] Feng Guan, Xiaoming Yu, Zeying Peng, Hongbo Xu, Yue Liu, Linhai Song, Xueqi Cheng1 at TREC-09. In Proceedings of the Eighteenth Text Retrieval Conference, Chinese Academy of Sciences, 2009.

[8] Nick Craswell, Dennis Fetterly, Marc Najork, Stephen Robertson, Emine Yilmaz. Microsoft Research at TREC 2009 Web and Relevance Feedback Tracks. Microsoft Research.[10] Steven Garcia, RMIT University at TREC 2009: Web Track, School of Computer Science and IT RMIT University, GPO Box 2476 Melbourne 3001, Australia.

[11] Zheng Ye1, Xiangji Huang1, Ben He1, Hongfei Lin. York University at TREC 2009: Relevance Feedback Track. Information Retrieval and Knowledge Managment Lab, York University, Toronto, Canada,

Information Retrieval Lab, Dalian University of Technology, Dalian, China.

[12] Rianne Kaptein, Marijn Koolen, Jaap Kamps. Result Diversity and Entity Ranking Experiments: Anchors, Links, Text and Wikipedia. Archives and Information Studies, Faculty of Humanities, University of Amsterdam, ISLA, Informatics Institute, University of Amsterdam.

[13] Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. In Proceedings of the seventh international World Wide Web conference (pp. 107–117). Amsterdam: Elsevier.

[14] McBryan, O. (1994). GENVL and WWWW: tools for taming the Web. In Proceedings of the first WWW conference. Amsterdam: Elsevier.

[15] Fujita, S. (2001). Reflections on aboutness TREC-9 evaluation experiments at Just system. In Proceedings of the ninth Text Retrieval Conference TREC-2001. NIST Special Publication #500-250.

[16] Singhal, A. & Kaszkiel, M. (2001). AT&T at TREC-9. In Proceedings of the ninth text retrieval conference TREC-2000. NIST Special Publication #500-249.

[17] Amati, G. & Carpineto, C. (2002). FUB at TREC-10 Web Track: A probabilistic framework for topic relevance term weighting. In Proceedings of the tenth text retrieval conference TREC-2001 (pp. 182–191). NIST Special Publication #500-250.

[18] Craswell, N. & Hawking, D. (2003). Overview of the TREC-2002 web track. In Proceedings of the eleventh text retrieval conference TREC-2002 (pp. 86–95). NIST Special Publication #500-251.

[19] Fujita, S. (2002). More reflections on aboutness TREC-2001 evaluation experiments at justsystem. In Proceedings of the tenth text retrieval conference TREC-2002 (pp. 331–338). NIST Special Publication #500-251.

[20] Hodong Li. “An Inverted Index Generator for CINDI”, Master’s Thesis, Computer Science, Concordia University, Canada, 2003.