Journal of Informetrics 7 (2013) 301–312

Contents lists available at SciVerse ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Review

Low-cost evaluation techniques for information retrieval systems: A review

Shiva Imani Moghadasi[a], Sri Devi Ravana[a,*], Sudharshan N. Raman[b]

[a] Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia
[b] Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, Malaysia

ARTICLE INFO

Article history:
Received 24 April 2012
Received in revised form
30 November 2012
Accepted 3 December 2012
Available online 10 January 2013

Keywords:
Retrieval evaluation
Effectiveness metrics
Relevance judgment
Test collection
Pooling

ABSTRACT

For a system-based information retrieval evaluation, test collection model still remains as a costly task. Producing relevance judgments is an expensive, time consuming task which has to be performed by human assessors. It is not viable to assess the relevancy of every single document in a corpus against each topic for a large collection. In an experimental-based environment, partial judgment on the basis of a pooling method is created to substitute a complete assessment of documents for relevancy. Due to the increasing number of documents, topics, and retrieval systems, the need to perform low-cost evaluations while obtaining reliable results is essential. Researchers are seeking techniques to reduce the costs of experimental IR evaluation process by the means of reducing the number of relevance judgments to be performed or even eliminating them while still obtaining reliable results. In this paper, various state-of-the-art approaches in performing low-cost retrieval evaluation are discussed under each of the following categories; selecting the best sets of documents to be judged; calculating evaluation measures, both, robust to incomplete judgments; statistical inference of evaluation metrics; inference of judgments on relevance, query selection; techniques to test the reliability of the evaluation and reusability of the constructed collections; and other alternative methods to pooling. This paper is intended to link the reader to the corpus of 'must read' papers in the area of low-cost evaluation of IR systems.
© 2012 Elsevier Ltd. All rights reserved.

Contents

* Corresponding author. Tel.: +60 3 7967 6348; fax: +60 3 7957 9249.
E-mail addresses: sdevi@um.edu.my, sdravana@gmail.com (S.D. Ravana).

1751-1577/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.joi.2012.12.001

1. Introduction

Performance characteristics of retrieval systems deals with the accuracy of produced results which are about how effective an information retrieval (IR) system is in retrieving relevant documents. In order to evaluate the effectiveness of IR systems, two different approaches that may complement each other can be adopted. These are: user-based and system-based methods. The user-based approach concentrates on observing the user's interactions with the system to quantify their satisfaction levels (Fidel, 1993). This method deals with obtaining and analyzing the user's feedbacks on the retrieval performance, user interface and other aspects of the system. The user-based method requires lots of human participation, and indeed will be costly and time consuming.

On the other hand, the system-based retrieval evaluation focuses on experiments that are aimed to evaluate the performance of the retrieval algorithm and considers users as an abstraction (Mandl, 2008). Such evaluation usually utilizes a test collection (Sanderson, 2010). The test collection consists of a document corpus, queries and set of judgments on relevance, which will be elaborated in Section 2.1 (Baeza-Yates & Riberio-Neto, 1999; Mandl, 2008; Melucci & Baeza-Yates, 2011). On the other hand, system scores are basically computed based on a chosen evaluation metrics. An evaluation metric quantifies the similarity between the set of documents retrieved by the systems (also known as *runs*) and a set of relevant documents (*qrels*) to see how good a retrieval system is. Due to repeatability, reusability and scalability characteristics of this system-based experiment, it provides a proper environment for evaluation and performing experimental experiences that makes this approach to be important (Baeza-Yates & Riberio-Neto, 1999).

Performing IR evaluation through the test collections is costly since part of this method relies on human effort. Assessing the relevancy of documents to a topic is a time consuming, and expensive task that has to be performed by human assessors who are usually specialists in one or more areas of knowledge. Due to the huge number of documents in the document corpus (to simulate the real world search engines), having a complete relevance judgment in a collection such as *TREC* is not viable. TREC is an initiative by the National Institute of Standards and Technology (NIST) and U.S. Department of Defence which provides the necessary common platform for research within the IR community for large-scale evaluation of retrieval methods. Table 1 shows the number and size of documents in the TREC document corpus, and number of topics in various TREC experiments from 1998 to 2010. For example, for TREC-2010 Web, to obtain a complete judgment set, a total of 1 billion documents should be assessed by the experts. Judging 1 billion documents with limited number of experts is nearly impossible as it will incur high cost in terms of judgment effort (hiring of suitable assessors to represent real users) and time-consumed during the judgment process. For example, by assuming that two documents could be successfully judged by an assessor within a minute (see Section 6.3, Sanderson, 2010), it will take about 347,000 days (or 950 years) to judge 1 billion documents. In order to judge more documents within a limited time, more assessors need to be hired which would also incur cost. This example shows that while the cost for relevance assessment could be easily quantified, there are other issues that need to be considered as well to ensure consistency in judgments by assessors and generate reliable evaluation results.

The *pooling method* was proposed by Spärk Jones and van Rijsbergen (1975), and adopted by subsequent IR evaluation initiatives in order to decrease the number of judgments that need to be created. In this method, a set of top d ranked documents returned by participating systems (in the TREC experiment) are selected to create the pool of documents that need to be judged. Then, all the duplicate documents are eliminated from the pool and followed by a judgment for relevancy by the assessors. The judgment set generated is called the *partial relevance judgments* because not all documents from the corpus were used for the assessment. All the documents outside the pool were considered as non-relevant. TREC was the first initiative that used partial relevance judgment based on a pooling method as a substitution for complete judgment (Spärk Jones & van Rijsbergen, 1975). Some other initiatives for experimental based evaluation such as Cross Language Evaluation Forum (CLEF), Forum for Information Retrieval Evaluation (FIRE), NACSIS Test Collection for Information Retrieval (NTCIR), Chinese Web Information Retrieval Forum (CWIRF), and the IR Initiative for Evaluation of XML retrieval (INEX) are also using similar methods to generate relevance judgment sets.

Using this partial relevance judgment set, participating systems can be fairly evaluated provided that the systems have contributed to the pool. Problem arises when new systems or improved systems need to be evaluated using this same relevance set. In the experimental based evaluation, reusing of the relevance judgments set is practiced because generating a new set of relevance judgments will incur additional time and effort. This practice of reusing judgments is difficult and

Table 1
Summary of TREC document collections and topics (Clarke, Craswell, & Soboroff, 2004; Clarke et al., 2009; Craswell & Hawking, 2002; Hawking, 1998, 2000; Xue et al., 2010).

Year	TREC experiment	Collection	Document size	No. of documents (million)	No. of topics
1998	TREC-7 VLC	VLC2	100 GB	18.5	50
2000	TREC-9 Web	VLC2, WT10G	100 GB, 10 GB	18.5, 1.69	50, 50
2002	TREC-2002 Web	.GOV	18 GB	1.25	50, 150
2004	TREC-2004 Web	.GOV	18 GB	1.25	225
2004	TREC-2004 Terabyte	.GOV2	426 GB	25	50
2009	TREC-2009 Web	ClueWeb09	25 TB	1000	50
2010	TREC-2010 Web	ClueWeb09	25 TB	1000	50

Download English Version:

https://daneshyari.com/en/article/523226

Download Persian Version:

https://daneshyari.com/article/523226

Daneshyari.com