Information Processing and Management 47 (2011) 18–36

Contents lists available at ScienceDirect

Information Processing and Management

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

Modeling information sources as integrals for effective and efficient source selection

Georgios Paltoglou [a,*], Michail Salampasis [b], Maria Satratzemi [a]

[a] University of Macedonia, Egnatias 156, 54006 Thessaloniki, Greece
[b] Alexander Technological Educational Institute of Thessaloniki, P.O. Box 141, 57400 Thessaloniki, Greece

ARTICLE INFO

Article history:
Received 31 July 2008
Received in revised form 11 February 2010
Accepted 12 February 2010
Available online 6 March 2010

Keywords:
Source selection
Distributed information retrieval
Federated search

ABSTRACT

In this paper, a new source selection algorithm for uncooperative distributed information retrieval environments is presented. The algorithm functions by modeling each information source as an integral, using the relevance score and the intra-collection position of its sampled documents in reference to a centralized sample index and selects the collections that cover the largest area in the rank-relevance space. Based on the above novel metric, the algorithm explicitly focuses on addressing the two goals of source selection; *high-recall*, which is important for source recommendation applications and *high-precision* which is important for distributed information retrieval, aiming to produce a high-precision final merged list.

For the latter goal in particular, the new approach steps away from the usual practice of DIR systems of explicitly declaring the number of collections that must be queried and instead focuses solely on the number of retrieved documents in the final merged list, dynamically calculating the number of collections that are selected and the number of documents requested from each. The algorithm is tested in a wide range of testbeds in both recall and precision-oriented settings and its effectiveness is found to be equal or better than other state-of-the-art algorithms.

1. Introduction

Distributed information retrieval (DIR) (Callan, 2000, chap. 5), also known as federated search (Si & Callan, 2003b), offers users the capability of simultaneously searching multiple remote information sources (i.e. search engines or specialized web sites) through a single interface. The importance of DIR has particularly augmented in recent years as the prohibitive size and rate of growth of the web (Lyman & Varian, 2003) make it impossible to be indexed completely. More importantly, a large number of web sites, collectively known as *invisible web* (Bergman, 2001; Raghavan & Garcia-Molina, 2001; Sherman, 2001) are either not reachable by search engines or do not allow their content to be indexed by them, offering their own search capabilities. Even publicly available, up-to-date and authoritative government information is often not indexable by search engines (Miller, 2007). Studies by Bergman (2001) have indicated that the size of the invisible web may be 2–50 times the size of the web reachable by search engines.

The distributed information retrieval process can be perceived as three separate but interleaved sub-processes: *source representation* (Callan & Connell, 2001; Si & Callan, 2003a), in which surrogates of the available remote collections are created. This process takes place before the user poses a query to the DIR system. *Source selection* (Callan, Lu, & Croft, 1995;

* Corresponding author.
E-mail addresses: gpalt@uom.gr (G. Paltoglou), cs1msa@it.teithe.gr (M. Salampasis), maya@uom.gr (M. Satratzemi).

0306-4573/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2010.02.004

Powell, French, Callan, Connell, & Viles, 2000; Si & Callan, 2003a), in which a subset of the available information sources is chosen to process the query, once it has been submitted and *results merging* (Craswell, Hawking, & Thistlewaite, 1999; Paltoglou, Salampasis, & Satratzemi, 2007; Si & Callan, 2003b), in which the separate results are combined into a single merged result list which is returned to the user.

This paper deals with the source selection problem. Previous research (Callan et al., 1995; Nottelmann & Fuhr, 2003a; Si & Callan, 2003a; Si, Jin, Callan, & Ogilvie, 2002) has shown that the source selection phase is vital to the overall effectiveness of the retrieval process. Research by Si and Callan (2004) has differentiated the source selection problem in two distinct but interrelated subproblems; the *high-recall* and *high-precision* goals.

The former deserves particular attention, since its definition is differentiated from the definition of recall for classical, centralized information retrieval environments. The high-recall goal for DIR environments is aimed at source recommendation applications, where the aim is to select a small set of the available resources that contain as many relevant documents as possible. A likely scenario of usage of such an application would be to recommend to users hidden web sites that are likely to contain significant amounts of information relevant to their information need, for browsing purposes. Recent work by Seo and Croft (2008) explicitly views the Blog Distillation task, under which the goal is to direct users to blogs that have a central and recurring interest in topic X, as a special case of source selection for source recommendation applications, achieving one of the best performances in the corresponding task in TREC 2007 (Macdonald, Ounis, & Soboroff, 2007). The high-recall task may in some ways be related to Broder's *navigational queries* under his taxonomy of web search (Broder, 2002), according to which a large percentage (approx. 68%) of users are mainly interested in reaching "a site that provides good information" on a specific topic. The above hypothesis remains an open research topic since no user-oriented studies have been done in federated search environments to verify it.

An alternative definition of *high-recall* for DIR environments is the elimination of collections that contain no relevant documents. The differences between the two definitions are subtle but could potentially lead to different approaches in dealing with the problem, from maximizing the number of relevant documents in recommended collections to minimizing the number of collections recommended that have no relevant documents. This definition may be closer to the definition of recall as it is applied in classic IR, but the initial given definition is the one that has been followed by most researchers and is the one that is adopted in this paper.

The *high-precision* goal aims at providing the user a final merged result list with the most relevant documents appearing in the top ranks and is related to classic IR precision. Although the two tasks are highly associated, they are implicitly distinct. Algorithms that perform adequately on one, are not guarantied to perform similarly on the other. The apparent discord is based on the ability (or lack of) of the remote collections to actually return their most relevant documents for merging, having been selected in the source selection phase. Therefore, for example, a collection with a significant number of relevant documents may not be able to retrieve them for merging, while another collection containing fewer relevant documents may retrieve most of them, thus significantly contributing to the quality of the final merged document list. The observation was originally made by Craswell (2000) and has been further explored in Si and Callan (2004), Si and Callan (2005) and Shokouhi (2007).

The source selection algorithm that is presented here explicitly focuses on both of the above goals, using a novel and simple metric for estimating the relevance of information sources by modeling each collection as a integral using the relevance score and the intra-collection rank of its sampled documents. In order to achieve the goal of *high-recall*, the algorithm selects the collections that cover the largest area in the rank-relevance space. For the *high-precision* goal, the algorithm divides the area covered by the remote collections into segments, each representing an estimation of the potential benefit of including their top-ranked documents in the final merged list and calculates the optimal distribution of retrieved documents in order to maximize the overall gain.

The rest of the paper is structured as follows. Section 2 reports on prior work. Section 3 describes the new methodology proposed in this paper. Section 4 describes the setup of the experiments conducted. Section 5 reports and discusses the results obtained and Section 6 concludes the paper, summarizing the findings.

2. Prior work

Source selection has received considerable attention in research the last years. In this section we present the most prominent work.

The STARTS initiative by Gravano, Chang, Garcia-Molina, and Paepcke (1997) is an attempt to facilitate the task of querying multiple document sources through a commonly agreed protocol. It provides a solution for acquiring resource descriptions in a cooperative environment, where remote collections provide important statistics about their contents.

When cooperation from collections is not available (i.e. isolated environments), techniques have been developed that allow for the estimation of their contents. Query-based sampling (Callan & Connell, 2001) is such a technique that creates collection samples through multiple one-term queries. Estimation of collection sizes is also possible through sample–resample (Si & Callan, 2003a) or capture–recapture (Shokouhi, Zobel, Scholer, & Tahaghoghi, 2006) methodologies.

GlOSS by Gravano, Garcia-Molina, and Tomasic (1999) is a source selection algorithm that uses document frequency and the sum of term weights within each remote collection. However, it is based on certain unrealistic assumptions about the distribution of terms and term weights within the documents (*high-correlation* and *disjoint* scenarios). CVV by Yuwono