International Journal of Information Management xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

International Journal of Information Management

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

Research Note

Enhancing information source selection using a genetic algorithm and social tagging

Fatma Zohra Lebib[a,b,*], Hakima Mellah[a], Habiba Drias[b]

[a] CERIST, Algiers, Algeria
[b] USTHB, LRIA, Algiers, Algeria

ARTICLE INFO

Keywords:
Information sources selection
Distributed information retrieval
Bio-inspired methods
Genetic algorithms
Social tagging

ABSTRACT

The selection of information sources in a distributed information retrieval environment remains a critical issue. In this context, it is known that a distributed information retrieval system consists of a huge number of sources. Ensuring retrieval effectiveness is to search only sources which are **likely to contain relevant information for a query**. An important number of heuristics exist among which we quote genetic algorithm that is used to solve the above problem. The proposed genetic algorithm consists in finding the best selection in large space of potential solutions; where a solution is represented as a combination of a set of sources. The improvement of selection accuracy is assured based on the user's track through the use of sources, to say that source description is enriched with tags from the tagging history.

1. Introduction

Even though the web can be seen as a single network of distributed repositories, many of traditional information retrieval approaches became difficult to beput into practice. One of the most important raison is the variety, heterogeneity and distributivity of information sources. The distributed information sources are asked for a query at the same time. This operation will certainly, return a huge information and consume a considerable time.

Distributed Information Retrieval (DIR) (Callan, 2000; Shokouhi & Si, 2011) provides a solution to the problem of searching on several dispersed information sources. The DIR system consists of three phases, namely *source description* (Callan & Connell 2001; Si & Callan, 2003), *source selection* (Callan, Lu, & Bruce Croft, 1995; Si & Callan, 2003a; Thomas & Shokouhi, 2009), and *result merging* (Craswell, Hawking, & Thistlewaite, 1999; Paltoglou et al., 2007; Shokouhi & Zobel, 2009; Si & Callan, 2003b). In the first phase, representations of available remote sources are created, containing important information about the sources such as their contents and their sizes. In the second phase, the DIR system selects a subset of information sources which are most useful for users' queries. The source description is used to estimate the relevance of each source, and to classify sources accordingly. Finally, the third phase combines documents retrieved from selected sources into a single ranked list before presenting the list to the end users.

The focus of this paper is the source selection problem in a context of a large number of sources. Previous research (Cetintas, Si, & Yuan, 2009; Nottelmann & Fuhr, 2003) showed that the source selection phase is vital for the overall effectiveness of the distributed research process.

In a multi-sources environment, users should look for several online information sources to satisfy their information needs. These sources are generally diverse and distributed and can be databases, web sites, academic research libraries, specialized shopping sites, search engines, etc. In such environment, sending the query to all sources is hard to handle insofar as some of them may not contain the desired information and generally results in decrease efficiency due to the introduction of irrelevant documents in the final results. Therefore, it is important to select a small number of sources from all available sources, which include as many relevant documents as possible to increase the research relevance in terms of recall and precision.

We formulized the problem of sources selection as a combinatorial optimization problem, which consists in finding the suitable combination (or selection) in a prohibitive search space containing all possible solutions (combinations). We search the solution, which maximizing the similarity between sources, composing a selection and the user query. We cope with this complex issue by the use of intelligent approaches, in particular the Genetic Algorithms (GAs) (Goldberg, 1989). GAs have been widely used by the scientific communities, not only for their simplicity of implementation but also for their effectiveness and efficiency when they are judiciously designed. GAs are considered robust and efficient (Eiben & Smith 2007) and outperform the analytical

* Corresponding author at: CERIST, Algiers, Algeria.
E-mail addresses: mfatma_zohra@yahoo.fr, zmatouk@mail.cerist.dz (F.Z. Lebib), hmellah@cerist.dz, hamellah@yahoo.fr (H. Mellah), h_drias@hotmail.fr (H. Drias).

http://dx.doi.org/10.1016/j.ijinfomgt.2017.07.011

methods for the large scale data (Drias, Khennak, & Boukhedra, 2009). They can be used to find good quality solutions to several spiny problems.

We consider that evolutionary algorithms (Bach, Fogel, & Michalewicz, 1997) are appropriate for source selection problem in distributed information retrieval for the following reasons (Bhatnagar & Pareek, 2012):

- The number of information sources that are increasing day by day makes the selection of a set of sources that may contain relevant information more complex and difficult to treat by analytical methods.
- The source selection problem can be considered as a search and optimization problem which aims to find near optimal sources from the available sources for a given query.
- In a large search space, it is important to explore and exploit each region of the search space to find the selection of sources with the highest relevance. The operators such as crossover and mutation perform such operations.

Source selection consists in finding the right sources to answer the query. It is based on source description and includes others techniques to evaluate the source relevance for the query. However, a description's accuracy is required for improving the selection effectiveness. In this work, in addition to the application of GAs for sources selection, and with the aim to improve the quality of source description, a social aspect is considered; it consists in exploiting the social tagging to enrich the sources description. With the social tagging additional information extracted from user's behaviors is integrated.

Social tagging applications (Mathes, 2004) allow users to freely associate labels also called keywords or tags, to resources. It is a common way of organizing resources for future navigation, filtering or search.

In general, users annotate items that are relevant for them, so the tags they provide can be assumed to describe their interests and needs. Moreover, it can be also assumed that the more a tag is used, the more important that tag is for the user. Analogously, tags assigned to items usually describe their contents. The more users annotate items with a particular tag, the better that tag describes the item contents.

These tags can identify the topics of bookmarked resources (topic(s) of an article), the tagger's opinion of the resources (opinion about a webpage) or for grouping task-related information (Example of performing tasks: to read, car search) (Golder & Huberman, 2006).

In our study, we consider tags that describe an item, where an item is a particular information source.

We encourage users to provide tags on the use of sources, in order to form a tags cloud per source. Through tags, users describe the sources with semantic annotations useful for source selection. The source tags can be exploited for additional information when selecting sources for a user query. The source tags can be matched using standard similarity measures against user requests.

2. The traditional metrics used for the source selection issue

Several traditional metrics exist for the information source selection. A rich and abundant literature exists for this subject. In the following, we present the most used by the information retrieval community.

The CORI algorithm (Callan et al., 1995) is one of the best-known DIR algorithms for collection selection, so it was used as a baseline in the research reported here. CORI adapts the INQUERY document scoring formula to score sources according to their vocabulary and document frequencies. This is commonly estimated based on sampled documents, but CORI treats each collection as compound "document" using document frequency instead term frequency.

The score P for each collection c, for query term t is given by:

$$P(t/c) = \Phi + (1 + \Phi).\ T.I \tag{1}$$

Where:

$$T = \frac{df_{t,c}}{df_{t,c} + 50 + 150 \times \frac{cw_c}{avg.cw}} \tag{2}$$

$$I = \frac{\log\left(\frac{N_c + 0.5}{cf_t}\right)}{\log(N_c + 1.0)} \tag{3}$$

$df_{t,c}$: The document frequency of t in c
cw_c: The number of words in the collection c
avg_cw: The average cw across all collections
N_c: The number of collections
cf_t: Collection frequency of t (the number of collections containing t)
Φ: The minimum belief component when t is available in c (default value is 0.4)

The belief P (q/c) is used by the CORI algorithm to classify the collections. The computation of P (q/c) consists in using the mean value of the beliefs of all the terms of the query.

As the aim of this study focusses on designing an intelligent approach with a specific modeling, we used CORI as a metric for the fitness function to optimize. Nevertheless, we can exploit any of the following metrics that will be described in the next section:

- Glossary of Servers Server (GlOSS) proposed by Gravano, Garcia-Molina, and Tomasic (1994).
- Cue Validity Variance (CVV) suggested by Yuwono and Lee (1997).
- KL (Kullback-Leibler) divergence presented by Xu and Croft (1999).
- Decision-Theoretic framework (DTF) proposed by Fuhr (1999).
- ReDDE (Relevant Document Distribution Estimation) developed by Si and Callan (2003a).
- UUM (Unified Utility Maximization) designed by the Si and Callan (2004).
- CRCS and CRCS conceived by Shokouhi (2007).

3. Related work

Several classic approaches were developed for the issue of information source selection. A rich and abundant literature exists for this topic. We present in the following, some of the main efforts deployed for this concern.

3.1. The classic sources selection approaches in a distributed environment

The first generation of source selection approaches, also known as big document approaches, represents each source as a concatenation of its documents. The big documents obtained are classified according to their lexical similarity with the query using standard information retrieval techniques based on tf (term frequency) and idf (inverse document frequency). Where, df (document frequency) is used instead of tf and icf (inverse collection frequency) instead of idf. The most well-known approaches are CORI (Callan, 2000; Callan et al., 1995), gGlOSS (Gravano, Ipeirotis, & Sahami, 1999) and CVV (Yuwono & Lee 1997). As prior research on different datasets has shown the CORI algorithm to be the most stable and effective of the three algorithms (Callan, Powell, French, & Connell, 2000; Powell & French 2003), we use it as a baseline algorithm in this work.

The second generation or small document approaches use a centralized index of sampled documents from different sources. The information sources are selected based on the ranking of their documents for a given query. The relevance of documents in sources is estimated to classify sources according to the number and position of their documents in a centralized ranking. Examples of these approaches are ReDDE (Si & Callan, 2003a), CRCS (Shokouhi, 2007) and others (Markov, Azzopardi, & Crestani, 2013; Paltoglou, Salampasis, &