Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Processing and Management 42 (2006) 1137–1150
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Testing the cluster hypothesis in distributed information retrieval

Fabio Crestani [a,*], Shengli Wu [b]

[a] Department of Computer and Information Sciences, University of Strathclyde, Glasgow G1 1XH, UK
[b] School of Computing and Mathematics, University of Ulster, Belfast, UK

Received 24 July 2005; received in revised form 5 December 2005; accepted 5 December 2005
Available online 31 January 2006

Abstract

How to merge and organise query results retrieved from different resources is one of the key issues in distributed information retrieval. Some previous research and experiments suggest that cluster-based document browsing is more effective than a single merged list. Cluster-based retrieval results presentation is based on the cluster hypothesis, which states that documents that cluster together have a similar relevance to a given query. However, while this hypothesis has been demonstrated to hold in classical information retrieval environments, it has never been fully tested in heterogeneous distributed information retrieval environments. Heterogeneous document representations, the presence of document duplicates, and disparate qualities of retrieval results, are major features of an heterogeneous distributed information retrieval environment that might disrupt the effectiveness of the cluster hypothesis. In this paper we report on an experimental investigation into the validity and effectiveness of the cluster hypothesis in highly heterogeneous distributed information retrieval environments. The results show that although clustering is affected by different retrieval results representations and quality, the cluster hypothesis still holds and that generating hierarchical clusters in highly heterogeneous distributed information retrieval environments is still a very effective way of presenting retrieval results to users.

Keywords: Information retrieval; Retrieval results presentation; Clustering; Experimental study; Distributed information retrieval

1. Introduction

With the ever expanding Web, users are faced with a huge number of information resources. How to quickly find what a user needs from this ''information ocean'' is a challenging issue. Although the single database/search engine solution seems to be efficient for that, it may have difficulties to collect all the information needed in practice, especially in relation to resources of the hidden Web. Distributed information retrieval

* Corresponding author. Tel.: +44 141 548 4303; fax: +44 141 548 4523.
E-mail address: f.crestani@cis.strath.ac.uk (F. Crestani).

0306-4573/$ - see front matter
doi:10.1016/j.ipm.2005.12.002

becomes an alternative and viable solution in these cases. In the MIND project,[1] that we have undertaken with four other partners from Europe and USA, we focused on two key issues of distributed information retrieval: resource selection and data fusion. When a user has routine access to a large number of multimedia digital libraries that are globally distributed, the first task he must undertake is selecting some of these resources to which to direct his queries. Considering the large amount of available resources (data sources, collections, Digital Libraries, etc.), this is a tedious work if done manually. So, setting up a broker to accomplish this resource selection task automatically is an attractive solution. The second task the system must undertake for the user is merging, organising and presenting to the user the results retrieved from all the selected resources. This task is called data fusion or results merging. This paper is concerned with this task.

There are several different ways of presenting the merged results to the user. The simplest solution is not to merge the results at all. Results are separately displayed according to their sources. Single-lists are often used for higher usability and effectiveness (Lee, 1997; Shaw & Fox, 1995; Vort & Cotterell, 1999). Another solution is to merge the results based on estimated document relevance. This approach has been investigated by many researchers and is widely used (see Section 2). However, one problem with this approach is that different resources might interpret the query in different ways and producing a single merged list of results is often affected by these different interpretations. Yet another solution is to cluster the results from the different sources based on retrieved document similarity. In this way different interpretations of the query might be captured. In addition, this enables users to browse effectively long results sets (Hearst & Pedersen, 1996). In fact, cluster-based retrieval results browsing is used by several Web search engines such as, for example, Northernlight[2] and Vivisimo.[3] Cluster-based retrieval results presentation is favourable to users since they can see an organised structure of the retrieved set of documents rather than a long list. For example, if a user submits a query with the word ''car'', the system might organise the retrieved documents in several different clusters, such as, for example, ''car rental'', ''auto'', ''buying'', ''sport'', ''car care'', ''classic car'', ''car audio'', etc. (example taken from Vivisimo), which may be helpful for the user to find the information he needs more quickly.

The effectiveness of retrieval results clustering lies in the cluster hypothesis that states that relevant documents tend to be more similar to each other than non-relevant documents (Jardin & van Rijsbergen, 1971). So according to this hypothesis, relevant documents should be found in the same cluster or clusters. It is obvious that the validity and effectiveness of the cluster hypothesis depends on the effectiveness of the clustering. Several hierarchical clustering models have proved to be effective for this purpose in traditional centralised information retrieval environments. However, the validity and effectiveness of the cluster hypothesis has not yet been investigated in distributed information retrieval environments.

Distributed information retrieval (DIR) is different from traditional centralised information retrieval (IR) in several ways:

- In DIR different resources often use different indexing and retrieval models.
- Different resources accessed by a DIR system often include different document collections and there may be overlap between the content of these collections.
- Different resources often have different result lists cut off policies, related to parameters like, for example, document access cost or bandwidth.

Consequently, with respect to clustering the results from different resources, these differences often carry the following consequences:

- The quality of the retrieved results from different resources often vary widely. This is caused by the quality of the search engines used in different resources and/or by the different number of relevant documents retrieved from each resource.

[1] More information about MIND can be found on the project web site at http://www.mind-project.org/.
[2] http://www.northernlight.com/.
[3] http://vivisimo.com/.