ABSTRACT

General purpose search engines, such as Google and Yahoo!, provide an easy mechanism for users to discover information on the Web. Despite their obvious advantages, they have a number of significant limitations, because they cannot reach or analyze a significant part of the information that is available.

Distributed Information Retrieval systems, employing collection fusion algorithms, offer a solution to the above problem, by allowing users to submit queries to multiple information sources simultaneously through a single interface, offering a much wider coverage of the available information.

This thesis deals with two of the main issues of designing and implementing efficient and effective Distributed Information Retrieval systems: *source selection* and *results merging*. The former deals with the ability of the system to select the most appropriate information sources to delegate the user query and the latter aims to produce the best possible final document list by merging to individual retrieved documents lists from the selected sources.

The new algorithms that are presented in this thesis are designed to function effectively in settings where information sources provide no cooperation at all, thus making them applicable in the widest possible set of environments and domains. The source selection algorithm that is put forth provides a novel modeling of information sources as regions in a space created by the documents that they contain. It provides a full theoretical framework for addressing the source selection problem, while at the same time effectively captures real-world observations and widely accepted notions in Information Retrieval. Extensive experiments demonstrate that it is able to obtain a performance that is at least as good as other state-of-the-art approaches and more often better.

The novel result merging algorithms that are presented are based on the supposition that search engines return only ranked lists of documents, without relevance scores, a scenario which is standard practice in current retrieval systems. They are both able to address the lack of information very effectively, demonstrating significant performance gains over

other state-of-the-art approaches. Additionally, the second algorithm unites the two general directions that the results merging problem has been approached in research, combining their advantages while minimizing their drawbacks.