Available online at www.sciencedirect.com
ScienceDirect
Information Systems 33 (2008) 658–680
www.elsevier.com/locate/infosys

# Joining the results of heterogeneous search engines

Daniele Braga, Alessandro Campi*, Stefano Ceri, Alessandro Raffio

*Dipartimento di Elettronica e Informazione, Politecnico di Milano, via Ponzio 34/5, Milan, Italy*

**Abstract**

In recent years, while search engines have become more and more powerful, several specialized search engines have been developed for different domains (e.g. library services, services dedicated to specific business sectors, geographic services, and so on). While such services beat generic search engines in their specific domain, they do not enable cross-references; therefore, they are of little use when queries require input from two or more of such services (e.g., "find papers in VLDB 2000 authored by a member of a specified department" or "books sold online written by prolific database authors" or "vegetarian restaurants in the surroundings of San Francisco"). In this paper, we study how to join heterogeneous search engines and get a unique answer that satisfies conjunctive queries, where each query can be routed to a specialized engine. The paper includes both the theoretical framework for stating such problem and the description of pragmatic solutions based on web service technology. We present several algorithms for the efficient computation of join results under several cost model assumptions.


*Keywords:* Search engines; Web services; Join strategies; Query optimization

## 1. Introduction

The current evolution of the web is characterized by an increasing number of search engines, ranging from generic ones (such as Google) to domain-specific ones (such as library search systems, or on-line recommendation systems for specific goods, or geo-localization services). While each search engine can be separately used to issue focused queries, their intrinsic limit is the inability to go beyond the purpose for which they have been developed; however, users are often interested in complex queries, ranging over multiple domains. Such queries can be only answered, at the current state of art, by a deep involvement of a knowledgeable user, who inspects search engines one at a time, feeding the results of one search as input to the next one. However, in an ideal scenario, users do not want to be bothered by distinctions between many searching systems and desire to have one common interface available for querying them; moreover, while they can accept a complex interaction when their query is rather complex, they certainly do not want to "cut-and-paste" query results into query inputs, as such approach is time consuming and yields imprecise results.

The focus of this paper is to develop techniques for merging the results extracted from two or more search engines. Our main idea is to design a system offering a common interface to several, known search services such that a user query can be

*Corresponding author. Tel.: +39 0223993644.
*E-mail addresses:* braga@elet.polimi.it (D. Braga), campi@elet.polimi.it (A. Campi), ceri@elet.polimi.it (S. Ceri), raffio@elet.polimi.it (A. Raffio).

0306-4379/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.is.2008.01.009

decomposed and its keywords can be routed to many of them in parallel. Given that each result is independently built, this approach consists then in *joining partial results into a composite result, to be presented to the user*. The recent availability of web services and XML as a means for generalized interoperability makes such an objective quite feasible.

Joining search engine results is indeed far more complex than joining conventional tables of relational databases: search engine results are ranked lists of complex XML structures which are returned in response to queries, and the pairwise join of the elements of such lists requires specialized optimization strategies. Join is indeed the right underlying paradigm, as we can talk about "join methods" which define the way in which web services are invoked and result lists are examined in order to produce the join results with the best overall performance.

In this paper, we do not address the orthogonal problem of how the user query can be dynamically decomposed and its keywords can be routed to the various search engines: we assume that users are provided with simple interfaces to predefined selections of search services which have been registered in our system, while the dynamic choice of the best specialized service for a given query is outside the scope of this research. Similarly, we do not address the general problem of integrating generic search engine results, which is an instance of the general problem of data integration [1–6], but we assume the existence of ad hoc coupling functions that build the result of joining two given service results by composing given XML elements extracted from them.

### 1.1. Relevant examples

Our ideal query involves several dimensions, each covered by a given search service. Examples of non-trivial queries that address orthogonal dimensions are the following:

- *Find a good vegetarian restaurant at approximately 30 miles from San Francisco.*
- *Find all VLDB authors from ETH Zurich.*
- *Find pairs of news from the Washington Post and the New York Times dealing with the same event.*

Any human actor recognizes at a glance that these queries can be considered as conjunctions of simpler queries over independent dimensions. The first one requires combining a geographic web service (say, MapPoint) and a recommending service expert in restaurants (say the Michelin Guide, or perhaps a more specific service restricting to vegetarian restaurants); the second one requires combining a web service dealing about publications (say, DBLP) and the staff of a given Faculty, as provided by a given wrapping service over a web site; the third one requires pairing services of two different web sources.

### 1.2. Technological scenario

The main technological innovation of the last years is the adoption of the service-oriented architecture (SOA); this technology allows applications to exchange information effectively and represents the turning point of several previous approaches to systems interoperability. Several services are already available on the web, not only for specific querying purposes, such as searching book catalogues or dynamically generating maps from geographic information systems, but also for querying general purpose search engines (such as Google) from within other applications. In addition, it is becoming possible to query those information sources which do not expose a web service interface by means of wrappers, which monitor the data contained in such sources and provide a service-based interface for querying the wrapped data [7,8]. This opportunity broadens the scope of the approach described in this paper and opens interesting scenarios for empowering web searches.

This paper discusses the issue of integrating search services so as to answer complex queries by leveraging those (simpler) search services, which are already available. More long-term research concerns web service composition (i.e., the ability to combine web services which are capable of performing tasks so as to build complex workflows) and Semantic Web Services [9,10] (i.e., the ability to select the services that best match a user's goal where both the goal and the services are semantically described by means of knowledge bases or ontologies). Compared to the above long-term research goals, the objective of effectively joining results from different search engines is much easier.

### 1.3. Approach

Let us consider again the problem of finding authors of VLDB papers of a department. This