Chapter 1: Search Computing

Stefano Ceri

Politecnico di Milano, Dipartimento di Elettronica ed Informazione,
V. Ponzio 34/5, 20133 Milano, Italy
stefano.ceri@polimi.it

Abstract. Search Computing is a new paradigm for composing search services. While state-of-art search systems answer generic or domain-specific queries, Search Computing enables answering questions via a constellation of dynamically selected, cooperating search services, which are correlated by means of join operations. The idea is simple, yet pervasive. New language and description paradigms are required for expressing queries and for connecting services. New user interfaces and protocols help capturing ranking preferences and enabling their refinement.

Keywords: Complex queries, multi-dimensional queries, search services, join operation, data integration, data visualization, process composition.

1 Beyond Page Search

Throughout the last decade, Internet search has been primarily performed by routing users towards the specific Web page that best answered their information needs. Major search engines, such as Google, Yahoo and Bing, crawl the Web and index Web pages, highlighting worldwide candidate “best” pages with excellent precision and recall; such ability has proven adequate to fulfill users’ needs, to the point that Web search is customarily performed by millions of users, both for work and leisure.

However, not all information needs can be satisfied by individual pages on the surface Web. On one hand, the so-called “deep Web” contains information which is perhaps more valuable than what can be crawled on the surface Web; on another side, as the users get confident in the use of search engines, their queries become more and more complex, to the point that their formulation goes beyond what can be expressed with a few keywords, their answers require more than a list of Web pages, and general-purpose search engines perform poorly upon them. According to search company’s experts, the number of complex queries that are not answered well by major search engines due to their intrinsic complexity is remarkably high and increasing. Many search interactions can be considered as part of a more complex process of expressing goals and achieving tasks, as discussed in the vision paper by Ricardo Baeza Yates (Chapter 2).

When a query addresses a specific domain (e.g., travels, music, shows, food, movies, health, and genetic diseases), domain-specific search engines do a better job
than general-purpose ones; but their expertise is focused upon a given domain. Thus, one can separately find best travel offers and interesting music shows, or conduct genetic analysis and investigate the related medical literature, but can hardly combine information from diverse yet related domains. An expert user can perform several independent searches and then manually combine the findings, but such procedure is cumbersome and error prone.

Search Computing aims at responding to **multi-domain queries**, i.e., queries over multiple semantic fields of interest, by helping users (or by substituting to them) in their ability to decompose queries and manually assemble complete results from partial answers; thus, Search Computing aims at filling the gap between generalized search systems, which are unable to find information spanning multiple topics, and domain-specific search systems, which cannot go beyond their domain limits.

Paradigmatic examples of Search Computing queries are: “Where can I attend an interesting scientific conference in my field and at the same time relax on a beautiful beach nearby?”, “Where is the theatre closest to my hotel, offering a high rank action movie and a near-by pizzeria?”, “Who are the strongest candidates in Europe for competing on software ideas?”, “Who is the best doctor who can cure insomnia in a nearby public hospital?”, “Which are the highest risk factors associated with the most prevalent diseases among the young population?” These examples show that Search Computing aims at covering a large and increasing spectrum of user’s queries, which structurally go beyond the capabilities of general-purpose search engines. These queries cannot be answered without capturing some of their semantics, which at minimum consists in understanding their underlying domains, in routing appropriate query subsets to each domain specific source and in combining answers from each expert to build a complete answer that is meaningful for the user.

## 2 State of the Art

Processing queries on multiple search engines is not new; **meta-search engines** are capable of routing the same query to multiple search engines and then presenting composite results. **Kosmix** is a new-generation meta-search engine connecting to over a thousand of sources by using their Web services. In **Kosmix**, the relevant data sources for a query are determined by matching the user’s keywords with a huge private concept taxonomy (of about a million nodes), after manually tagging the data sources with the same taxonomy concepts. **Kosmix**, then, routes the query to all data sources, without attempting source integration. Results are collected from **Google**, **Yahoo**, **Flicker**, **YouTube**, **Twitter**, and so on, and presented to users; sources typically include very popular search engines, but sometime also domain-specific sources, such as the **Day-Life** or **Slate** (in the news domain). While **Kosmix** has the ability of showing individual results from many distinct data sources, it doesn’t integrate multiple domains, and therefore cannot answer complex queries; rather, it can answer simple queries by retrieving data from a plurality of sources. Yet, **Kosmix** demonstrates that Web services are viable methods for getting information from remote sources.

**Vertical search engines** are focused upon a single domain, e.g., hotels (**Booking**) or flights (**Tuifly**), which are well-understood in terms of data quality and ranking