NEED FOR PROTOCOLS AND STANDARDS FOR A DECENTRALIZED WEB SEARCH PARADIGM

Sergio R. Coria

University of the Sierra Sur
Calle Guillermo Rojas Mijangos S/N, Col. Ciudad Universitaria
70800 Miahuatlan de P. Diaz, Oax., Mexico

ABSTRACT

This paper proposes to create a new paradigm on services for information search on the web, involving new protocols and standards. For more than one decade, the current paradigm has consisted in services that have been provided by a small number of private companies on an advertising business scheme, mainly. This concentration involves a number of risks and disadvantages to WWW users. One of the risks is the service vulnerability because of the large dependence on the infrastructure of one or two providers. Scalability might become into a concern because of the huge increasing of information on the WWW. The current paradigm also involves economical and political disadvantages because the providers decide which websites are allowed to be listed or not as well as their rankings; this can bias the search results. Therefore, this paper aims to suggest general lines for research and development for a new, global, non-for-profit, paradigm for crawling, indexing and searching information on the WWW.

1. INTRODUCTION

Since the 1990 decade, service for information search on the World Wide Web (SISW) has been provided by a small number of private companies on an advertising business basis. A general explanation on how SISW work can be found in (Manning *et al.*, 2008), and Brin and Page (1998) present a specific description of Google, the most influential SISW at the present time. A historical overview of infrastructure and algorithms for information search on WWW is presented by Evans *et al.* (2005).

The current paradigm presents a series of technological, economical and political concerns to cyberspace users. Concentration in a small number of providers involves, among others, the risk that the service is more vulnerable because of the large dependence on their infrastructure only. Information increasing on the WWW can impact on infrastructure scalability. A non-technological issue is the bias in both search results and rankings that can be introduced by the service providers. Spink *et al.* (2006) show how search results differ among the major providers, suggesting that bias can be involved.

Centralization of web contents indexes and focus on profit orientation seem to be two important weaknesses of the current paradigm. Thus, index decentralization and a non-for-profit approach are desirable characteristics of a future, improved, paradigm. Index decentralization needs to be supported by a large number of servers; therefore, new protocols and standards to communicate multiple locations can be useful. A non-for-profit approach needs publicly accessible hardware infrastructure to reduce index hosting costs.

2. PROTOCOLS AND STANDARDS

Web search depends on creating and updating indexes of the WWW contents. This is a highly complex task that needs efficient algorithms as well as infrastructure with high storage and speed capabilities. Brin and Page (1998) state that distributed indexing systems, *e.g.* GlOSS (Gravano, 1994), can be *the most efficient and elegant technical solution for indexing*. Furthermore, they consider that distributed indexing would

improve searching drastically. Other more recent studies about distributed indexing for the WWW are (Melnik *et al*., 2001) and (Baeza-Yates *et al*., 2009). Brin and Page comment the difficulty to use distributed indexes on a worldwide basis because of the high administration costs of setting up large numbers of installations. However, they also suggest that this can become feasible by reducing the costs.

Our claim is that low-cost distributed indexing can be implemented as an inherent functionality of web servers if protocols and standards for crawling, indexing and ranking web contents are created. Once these are available, free software solutions can be developed to automatically create and manage local indexes. If every web server has an index and a ranking list of its local contents, the web search process can be easierly performed by a larger number of search engine providers, either commercial or non-commercial. This way, openness and auditability can be characteristics of the new paradigm, and bias in web searching can be reduced or even eliminated.

The definition of protocols and standards for crawling, indexing, ranking and searching can be started by enumerating features (*i.e*. functionalities) to be satisfied by web server software and by search engines. A series of useful features are suggested in sections 5 and 6.

3. INFRASTRUCTURE

In the new paradigm, a bottom-up approach can be considered, *i.e*. a fully- or semi- distributed searching service that operates on a set of publicly available servers and indexes can be implemented. No new infrastructure is needed because typical web servers can be used as hosts for their own contents indexes by using the new standards and protocols. This way, web hosting providers can allocate both forward and inverted indexes along with their hit lists.

The new providers of searching services would need less hardware resources for their search engines, and their indexes can be construed by inverted indexes and ranking tables only which can be produced from those located at web hosting servers. A probabilistic approach for distributed indexing can be adapted on a basis similar, for instance, to (Gravano *et al*. 1994). A fully non-for-profit paradigm can be implemented by defining a hierarchical structure and standardized protocols on a basis inspired on the DNS server global infrastructure. General lines to organize a distributed indexing scheme are suggested in section 4.
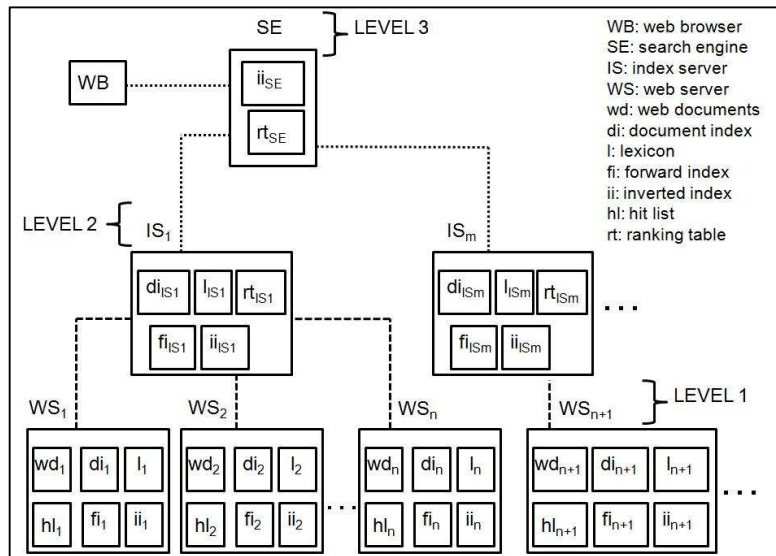

Figure 1. Basic architecture of the new paradigm

4. HIERARCHICAL STRUCTURE OF DISTRIBUTED INDEXES

In the current paradigm, search engines (*SE*) create forward indexes (*fi*), inverted indexes (*ii*) and other data structures, but no index is created at web servers (*WS*) locally. In the new paradigm, every *WS* creates a set of local indexes. In addition, index servers (*IS*) are proposed to allocate mid-hierarchy indexes that subsume data from those located at a set of *WS*. Figure 1 presents a hierarchical index structure for the new paradigm.

Three index levels are needed in the new paradigm:

Level 1: local indexes and hit lists that are located at every *WS*, the lowest level of the hierarchy.

Level 2: intermediate indexes and ranking tables that are located at every *IS*. This level can be construed by a number of sublevels.

Level 3: general indexes and ranking tables that are located at search engines (*SE*).

Structures at level 2 contain data that subsume data from level 1. *IS* can be hosted by web hosting providers. Level 3 contains data subsumed from level 2. Assignation of *WS* to *IS* needs to be ruled by a standard or protocol to guarantee the following conditions:

i.   At least one *IS* is assigned to every *WS*.

ii.   At most *n WS* are assigned to every *IS*.   *n* can be defined for each particular *IS* in accordance with a set of *IS* classes that depend on infrastructure capabilities. These involve storage size, hardware speed and networking bandwidth. Smallest *IS* can constitute, for instance, *class A*, other larger can constitute *class B*; other even larger, *class C*, etc. This way, *n* can be, for instance, 2 for *class A*; 4 for class B, 8 for class C, etc.

iii.   Every *IS* is available to every *SE*.

5. NEW FEATURES IN WEB SERVER SOFTWARE

Two of the most important new features to be available and standardized in web server software are indexing and ranking. Both involve data structure and algorithm issues that might be addressed on a basis similar to (Brin and Page, 1998) and (Melnik *et al*., 2001), as described below.

5.1 Indexing

Standardized data structures (see figure 1) for indexing local contents at every web server (*WS*) should store:

i.   A document index (*di*).

ii.   A lexicon (*l*).

iii.   A forward index (*fi*).

iv.   An inverted index (*ii*).

Similar data structures are needed at index servers (see *IS* in figure 1) to store data on a subsumed basis. A number of data structures that are present at level 1 might not be necessary at levels 2 or 3.The subsuming process can be based on a probabilistic approach, similar to (Gravano *et al*. 1994). Standardized indexing algorithms should:

i.   Create and update the data structures enumerated above.

ii.   Minimize data transfer among different levels of the architecture.

5.2 Ranking

Standardized data structures (see figure 1) for ranking should store:

i.   A hit list (*hl*) that is created from only local information and is hosted at every *WS* (level 1).

ii.   A ranking table (*rt*) that is created from data available at a number of *hl* and other additional data (*e.g*. link graphs). An *rt* is hosted at every *IS* (level 2).

Standardized ranking algorithms should:

i.   Create and update these data structures.

ii.   Minimize data transfer among different levels of the architecture.

6. FEATURES IN WEB SEARCH ENGINES

Since a series of indexes, hit lists and ranking tables can be available at *IS* (and also at *WS*), a *SE* does not need to create them from scratch. Instead, *SE* can take advantage of them. Therefore, *SE* should:

i.   Store and create data structures at level 3 that subsume data from level 2 and, maybe exceptionally, from level 1.
ii.   Update these data structures.
iii.   Minimize data transfer among different levels.

Crawling in the new paradigm is significantly different from that in the current. Even more, crawling could be unnecessary. The main reason is that levels 1 and 2 are responsible for creating and updating their indexes, hit lists and ranking tables on a bottom-up basis. Therefore, *SE* does not need to access *WS* directly. Rather, crawling is replaced by a new, different, process in which *SE* uses data structures of *IS* at level 2.

If standards for level 1 are established, other alternate architectures can be implemented for the new paradigm using other different data structures and features at levels 2 and 3. Regarding *WB*, new features could be incorporated, so that *WB* can perform a smarter role in information search and retrieval. Nevertheless, the increasing in network traffic that could be produced by smarter *WB* is a constraint to be taken into account.

7. CONCLUSIONS AND FUTURE RESEARCH

This paper has addressed the need for standards to crawl, index and rank web contents in order to provide web server software with local indexing features that support a global, distributed, indexing scheme. These standards and protocols can become guidelines to create free software applications so that every web server can automatically create and manage local indexes. If these functionalities become inherent to WWW protocols and servers, dependence on centralized search services can be reduced. Future research should address an evaluation of the best algorithms and data representations for these purposes in order to propose preliminary standards.

REFERENCES

Baeza-Yates, R. *et al*., 2009. On the Feasibility of Multi-site Web Search Engines. *Proceedings of ACM 18th Conference on Information and Knowledge Management (CIKM 09)*. Hong Kong, China, pp. 425-434.

Brin, S. and Page, L. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Proceedings of the Seventh International World-Wide Web Conference (WWW 1998)*. Brisbane, Australia.

Evans, M.P. *et al*., 2005. Search Adaptations and the Challenges of the Web. *IEEE Internet Computing*, Vol. 9, Issue 3, pp. 19-26.

Gravano *et al*. 1994. The Effectiveness of GlOSS for the Text-Database Discovery Problem. *Proceedings of the 1994 ACM SIGMOD International Conference On Management Of Data*. Minneapolis, USA, pp. 126-137.

Manning, C. D. *et al*. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, USA.

Melnik, S. *et al*. 2001. Building a Distributed Full-Text Index for the Web. Building a Distributed Full-Text Index for the Web. *ACM Transactions on Information Systems*, Vol. 19, No. 3, July 2001, Pages 217–241.

Spink, A. *et al*. 2006. Overlap among Major Web Search Engines. *Proceedings of the Third International Conference on Information Technology: New Generations (ITNG'06)*. Las Vegas, USA. pp. 370-374.