

DIARY: Distributed Architecture for Information Retrieval Techniques on Internet TIC2001-0547

Carmen Guerrero, Víctor Carneiro, Fernando Bellas, Angel Viña,
Fidel Casheda, Alberto Pan, Manuel Álvarez, Juan Raposo
Dpto. de Tecnoloxías da Información e as Comunicacóns
Universidade de A Coruña
e-mail: {clopez, viccar, fbellas, avc, fidel, apan, mad, jrs}@udc.es

Abstract

The final objective of this project is to realize a leading technological development in the Internet information retrieval area. Specifically, we are going to study the two main techniques of Internet information retrieval: directories and search engines. In the first phase we have examined the Web directories, with the objective of developing a new data architecture model to improve the performance of these systems. Next, and using the previous experiences, the project is focused on the search engines. In this phase, we are studying several distributed architectures for search engines, taking into account the new restrictions of the World Wide Web (large volume of information, distributed data, volatile data).

Keywords: information retrieval technology, distributed, Internet, robot, Web directory, search engine

1 Project objectives

This project is focused on the Information Retrieval (IR) technologies used in the World Wide Web, taking into account the two main types of IR systems: Web directories and search engines. The first phase of the project is related with the Web directories, studying different data architectures to improve the performance. And the second phase of this project deals with the search engines and different distributed architectures to handle the high amounts of information available nowadays.

The main objective in the first phase of the project is the improvement in the performance of the Web directories. With this purpose, a set of tools should be designed, and the analysis of different data architectures would be considered.

The objectives in the second phase of the project are more general, considering from a global point of view the search engines and researching its two main parts: the robot and the search engine itself. The most important restriction on the actual search engines is the high amount of information to be indexed and searched. In this way, the project will analyse the distribution of the whole architecture of the system: the robots and the data structures, in order to improve the

response times for the users' queries and to improve also the download and indexing time of the documents.

This project was scheduled to last three years, and each phase was designed to comprise half the time of the whole project. At the present moment the project is in the beginning of the second phase, with the first phase already concluded with some important results, represented as multiple publications in national and international conferences and two theses presented by two researchers of this project.

2 Success achieved in the project

Information retrieval systems appear in the Web with the purpose of managing, retrieving and filtering the information available in the WWW. There are three basic tools to locate information in the Web: search engines, Web directories and meta-searchers. Search engines index, ideally, the whole documents available in the Web, placing quantity above quality of its contents. Web directories are an ontology of the Web. The most relevant documents are classified according to topic, placing quality above quantity of documents. Finally, meta-searchers just resend queries to other search systems, and later they reorder the results.

In this project we analyse in detail different aspects of the Web directories and the search engines, as the most relevant IR systems available nowadays on the Web. This section is divided as follows: the first subsection examines the first phase of the project focused on the Web directories, whereas the second part describes the initial results obtained and the future works related with the research associated with the search engines, in the second phase of this project.

2.1 First phase: Web directories

Web directories are based on a hierarchical structure of categories, where the documents are classified. This structure is defined as a directed acyclic graph of categories, since one node may present several parents. Equally, one document may be catalogued in different categories. This provides Web directories with a great power and cataloguing flexibility.

Several improvement areas have been defined in the Web directory systems. Initially, some simple management tools were developed to help in the cataloguing process associated with any Web directory. Basically, these tools will discover the most relevant categories associated with a new document and will estimate the division of a category into two if the number of associated documents was too high.

Next, an exhaustive statistical analysis of the users of a Web directory was performed, using the information provided by the log of a real Web directory. The first objective of this analysis was to confirm the main conclusion from previous works: Web users differ significantly from traditional Information Retrieval users. Mainly in the following aspects: Web users will use very few words per query (approximately 2.37) and they will check very few screen results (usually no more than 2). And moreover, Web users do not usually employ logic operators performing very simple and basic queries.

The second part of this analysis obtained very useful information to understand the behaviour of the Web users as a group when they are using a search engine. Specifically, we deduced that the number of searches, the number of documents and categories viewed per minute follow a Poisson

distribution, with a variable mean through time. But, the most important conclusion is that there is a relationship among the number of searches, documents viewed and categories visited, which was estimated in this research.

This statistical analysis is the basis for the design and development of a simulation tool used in the performance evaluation of Web IR systems, named *U_{sim}*. This simulation tool contributes in the performance evaluation process in two different ways: estimating the saturation threshold of the system and in the comparison of different search algorithms or engines. The latter point is the most interesting because the comparison using different workload environments will achieve more accurate results (avoiding erroneous conclusions derived from ideal environments). From a general point of view, *U_{sim}* intends to be an approximation to some new performance evaluation techniques specifically developed for the Internet search engines.

However, the main research was focused on the data structures used in a Web directory. Nowadays, the inverted index is the indexing data structure that provides a better response time and storage space balance. Although, when using complex hierarchical data structures its performance could be not optimal.

In the case of the Web directories, this is evident in a special type of queries named *restricted queries*. In this case, the search is restricted to those documents associated with an area of the ontology specified by the root node to which the user has navigated. The standard search process is based on an inverted file structure, which relates keywords to their associated documents, while the category browsing process is based on an inverted file, which associates each category to its Web documents. On the contrary, the search process restricted to one area of the categories graph must combine both results lists in an efficient way.

In our work we have improved the performance of the restricted searches up to 50% using an hybrid data structure, based on inverted files with multiple signature files (using superimposing codes) embedded.

Initially, the superimposing code technique and the signature files were adapted to the environment of a directed acyclic graph of categories. Basically, a unique signature is associated with each category, and each documents calculates a superimposed code superimposing the signatures of its associated categories. A key aspect is the determination of the system parameters that directly influence the performance of the signature files. Only the increase in the percentage of documents associated to several categories and the percentage of categories with several parent nodes have a negative repercussion. Besides, these parameters have the feature of remaining stable throughout the life of a Web directory.

On the contrary, parameters such as the number of documents and categories of the search system that tend to increase with the time do not affect the false drop probability.

Therefore, the superimposing codes used in the signature file technique adapt perfectly well to a Web directory and its categories graph environment, obtaining a stable performance in the dynamic context of these systems.

To dynamically generate the signature file associated with each query the signatures files are inserted into each inverted list of the inverted file. In this way, a hybrid data structure is defined that will maintain the advantages of both data models: the signature files for the restricted queries and the inverted files for the standard queries.

With this hybrid architecture for the Web directories, two variants are defined: the hybrid model with total information and the hybrid model with partial information. The former corresponds to the direct application of the superimposing codes technique to the categories graph. In this case,

each and every one of the categories have an associated signature. In the latter the superimposing codes will be applied only to the first levels categories, whereas the remaining categories will inherit the signatures of the higher levels.

The main disadvantage of this hybrid architecture is the increase in the storage space required. In fact, the variant with total information will nearly double the space required by the inverted file used in the standard searches. But, the hybrid model with partial information was specially designed and developed to reduce the storage space requirements, requiring nearly 50% less disk space than the other variant.

For the performance evaluation of this new architecture model the simulation tool previously developed was used to generate different workload levels and so determine more accurately the performance of our hybrid models versus a basic model.

The results show that the variant with total information outperforms by 50% the basic model on a void, load or medium loaded environment, but in a high workload situation its response times are degraded, being equivalent to the basic model. On the other hand, the variant with partial information improves by 50% the performance of the basic model in all workload situations. The worse performance of the total information model is due to the excessive size of the index of keywords and documents. A high load situation implies a high number of searches reaching the index, so the disk undergoes more reading operations.

In the hybrid architecture proposed, the inexact filtering application allows a reduction of the response times. However, in the total information variant, the time required by the search process minimises the positive effect of the inexact filtering. On the contrary, in the hybrid model with partial information, the search time is inferior (since it requires less access to the disk), adding to the benefits of the inexact filtering, which allows an improvement in response timing of up to 50% with regard to the basic model.

Also, the implementations carried out have proved to be flexible enough with regard to the number of documents that the system can support, and also with regard to the number of categories in the directory.

From a global point of view the research carried out is a first step in the study of the data structures associated with the Web directories, and in some future works some other data structures will be examined.

2.2 Second phase: search engines

The second phase of the project is being developed at the moment and is focused on the search engines. The ideal objective of a search engine is to index the whole Web. However there are some special characteristics of the World Wide Web that affect directly to the search engines:

- Large volume: the exponential growth of the Web introduces scaling issues that are difficult to cope with.
- Volatile data: due to Internet dynamics, new computers and data can be added or removed easily.
- Distributed data: due to the intrinsic nature of the Web data spans over many computers and platforms.

The characteristics of the Web imply that the search engines will have to index a incredibly high amount of information and to update the frequent changes on the documents. This leads to the analysis of the data structures used in this systems.

Associated with this research area of the project, one member of the research group (Dr. Fidel Cacheda) has cooperated with the IR research group of the University of Glasgow, with a stay of six months. During this stay, a deeper analysis of the data structures used in the search engines was developed and a new research subject was introduced: the Combination of Evidences.

The combination of evidences is based on the idea that different queries require an alternate combination of content analysis (from traditional IR) and link analysis (i.e. PageRank or HITS). Experience on the Web suggests that queries on very specific topics can hardly benefit from link analysis; and an increased precision among the top ranked documents is observed for queries on a broad topic, when link analysis is performed.

As a consequence, there is a need for optimal combination of results obtained from content and link analysis with respect to a query. Nowadays, this combination is static and independent of the queries. With our work we aim to introduce a dynamic combination of evidence depending on each query. Intuitively, the contribution of link analysis should be higher for generic queries and lower for specific ones.

In our work we have identified several parameters (associated with a query) that could determine the generality or specificity of a query. And so, these parameters will establish the type of combination between the content and link analysis results.

Although this work is on an early stage, in cooperation with IR group of the University of Glasgow we have presented a proposal to the TREC 2003 (Text REtrieval Conference), one of the major events in the IR research, and specifically in the Web IR research. The results on the competition with other research groups have not been published, although some motivating conclusions are expected.

On the other side, the benefits of a distributed architecture applied to the search engines data structures was also examined in cooperation with the IR research group of the University of Glasgow. The use of distributed information retrieval systems is the only way to cope with the continuously increasing number of documents to be indexed in many environments (Web, intranets, digital libraries) and the limitations of a single centralized index (lack of scalability, server overloading and failures). In our work we have analysed the effectiveness of a distributed, replicated and cluster architecture for a distributed information retrieval system simulating a massive cluster of workstations (up to 4096). A collection of 94 million documents and 1 TB of text is used to test the performance of the systems.

The main results prove that in a purely distributed information retrieval system, the brokers become the bottleneck due to the high number of local answer sets to be sorted. In a replicated system, the network is the bottleneck due to the high number of query servers and the continuous data interchange with the brokers. A cluster system will outperform a replicated system if a high number of query servers is used, due mainly to the reduction of the network load. Although a change in the distribution of the users' queries could reduce the performance of a cluster system.

3 Results indicators

Two thesis have been presented by two members of our research group directly related with this project. The thesis referenced at [1] concentrates the main conclusions obtained from the first part of this project. Also, several national and international publications expose the results obtained in this part. Some of the most relevant papers are the ones referenced at [4], [5] and [6], where the

hybrid architecture is described and the improvement in the performance is exposed (some other articles have been published on this subject, although are not included in this document). The paper described in [7] exposes the methodology used in the performance evaluation of our models, and in general, in any Web IR system.

Another important point is the cooperation with the IR research group of the University of Glasgow, led by the professor J.C. van Rijsbergen. This research group is one of the main institutions in the IR area on the world. This cooperation was done through the stay of Dr. Fidel Cacheda during six months with the IR group, working directly with professor Iadh Ounis, leader of the Web IR area. In cooperation with them two main research studies are being developed: the combination of evidence and the distributed architectures for Web IR systems.

The article referenced at [3] just describes the initial conclusions obtained in the combination of evidence subject. As mentioned before, this work was presented to the TREC 2003 competition, and by the moment the results have not been published yet.

With regard to the distributed architectures, some publications will soon be proposed to some journals and international conferences.

Some new research areas are being investigated directly related with this project. First of all, the thesis referenced at [2] is associated with the Information Retrieval of structured and semi-structured information. A completely functional mediator system developed during this work is described in [9] and the WARGO system for web wrapper generation is described in [10]. Directly associated with the Internet search engines is the portal framework designed and developed, described in [8]. And associated with the important network restrictions of the distributed IR systems, new network technologies are being researched, especially the active networks [11].

4 References

- [1] Tesis doctoral “*Arquitectura de Datos Avanzada de un Directorio Web, con Optimización de Consultas Restringidas a una Zona del Grafo de Categorías*”. Fidel Cacheda Seijo, director: Angel Viña Castiñeiras. 2002.
- [2] Tesis doctoral “*Un Sistema Mediador para la Integración de Datos Estructurados y Semi-estructurados*”. Alberto Pan Bermúdez, director: Angel Viña Castiñeiras. 2002.
- [3] F. Cacheda, A. Viña, “*Combinación de Evidencias para la Recuperación de Información en el Web: Análisis del Ámbito de las Consultas*”. En Actas de Jornadas de Ingeniería Telemática 2003 (JITEL 2003).
- [4] F. Cacheda, V. Carneiro, C. Guerrero, A. Viña, “*Optimization of Restricted Searches in Web Directories Using Hybrid Data Structures*”. En Proceedings of the 25th European Conference on IR Research, ECIR 2003. ISBN: 3-540-01274-5. Lecture Notes in Computer Science 2633, pp: 436-451. Springer-Verlag, 2003.
- [5] F. Cacheda, A. Viña, “*Inverted files and dynamic signature files for optimisation of Web directories*”. En Proceedings of the 11th World Wide Web Conference (WWW2002), ISBN: 1-880672-20-0. ACM Press, 2002.
- [6] F. Cacheda, A. Viña, “*Optimización de Directorios Web mediante Estructuras de Datos Híbridas*”. En Actas de I Jornadas de Tratamiento y Recuperación de Información, ISBN: 84-9705-199-8. Editorial de la UPV, 2002.

- [7] F. Cacheda, A. Viña, "Performance evaluation of Web Information Retrieval Systems and its application to e-Business". En Proceedings of the e-2002 (e-Business and e-Work Conference and Exhibition), ISBNs: 1-58603-284-4 (IOS Press) y 4-274-90541-1-C3055 (Ohmsha). Vol. 1, pp: 725-732. IOS Press, 2002.
- [8] F. Bellas, D. Fernández, I. Toral, A. Muiño, "Towards a Generic and Adaptable J2EE-based Framework for Engineering Personalizable "My" Portals", Proceedings of the IADIS International Conference WWW/Internet 2003, Algarve, Portugal, 5-8 November 2003.
- [9] A. Pan, M. Álvarez, J. Raposo, P. Montoto, V. Orjales, A. Molano, A. Viña. "Mediator Systems in E-Commerce Applications". Proceedings of the 4th IEEE International Conference on Electronic Commerce and Web Information Systems (WECWIS). IEEE Computer Society Press. ISBN: 0-7695-1567-3, 0-7695-1568-1, 0-7695-1569-X pp 228-235 (2002)
- [10] A. Pan, J. Raposo, M. Álvarez, J. Hidalgo, A. Viña. "Semi-Automatic Wrapper Generation for Commercial Web Sources". Proceedings of the IFIP WG8.1 Conference on Engineering Information Systems on the Internet Context (EISIC). Kluwer. ISBN: 1-4020-7217-1, pp 265-283 (2002)
- [11] Puentes, F., Carneiro V., "Active Node Implementation in the context of a virtual active network orientated to collaborative environments", CE: The vision for the Future Generation in Research and Applications, Vol. I (p.p. 1199-1204), July 26-30, 2003, Madeira, Portugal.