Inverted files and dynamic signature files for optimisation of Web directories

Fidel Cacheda, Angel Viña

Department of Information and Communication Technologies

Facultad de Informática, University of A Coruña

Campus de Elviña s/n, 15071 A CORUÑA, SPAIN

Telephone: +34-981-167000 Fax: +34-981-167160

Email: {fidel, avc}@udc.es

ABSTRACT

Web directories are taxonomies for the classification of Web documents. This kind of IR systems present a specific type of search where the document collection is restricted to one area of the category graph. This paper introduces a specific data architecture for Web directories which improves the performance of restricted searches. That architecture is based on a hybrid data structure composed of an inverted file with multiple embedded signature files. Two variants based on the proposed model are presented: hybrid architecture with total information and hybrid architecture with partial information. The validity of this architecture has been analysed by means of developing both variants to be compared with a basic model. The performance of the restricted queries was clearly improved, specially the hybrid model with partial information, which yielded a positive response under any load of the search system.

Keywords

Web directory, data architecture, hybrid data structure, signature files, inverted files.

1 INTRODUCTION

IR systems appear in the Web with the purpose of managing, retrieving and filtering the information available in that huge data base constituted by the WWW.

There are three basic ways to locate information in the Web: search engines, Web directories and meta-searchers [1]. Web directories are an ontology of the Web, and have the added value of possessing a search process combined with the navigation one, which improves the quality of the obtained results. In this case, the search is restricted to those documents linked to an area of the ontology specified by the root node to which the user has navigated.

This paper examines in detail the restricted searches characteristic of Web directories, from the point of view of the data structures used and the performance obtained. Performance is improved by using a hybrid data model composed of an inverted file and dynamic signature files.

2 BASIC MODEL

A Web directory consists of three basic components. The vocabulary stands for the key words indexed both in the documents and in the directory categories. There is a structure which represents the hierarchy of categories existing in the directory, typically composed of a directed acyclic graph. A small document file is required with the basic information about each of them (URL, title and description). For representing the relationships among these three items, the inverted file structure constitutes at present the index technique which has the best performance for huge data volumes [1].

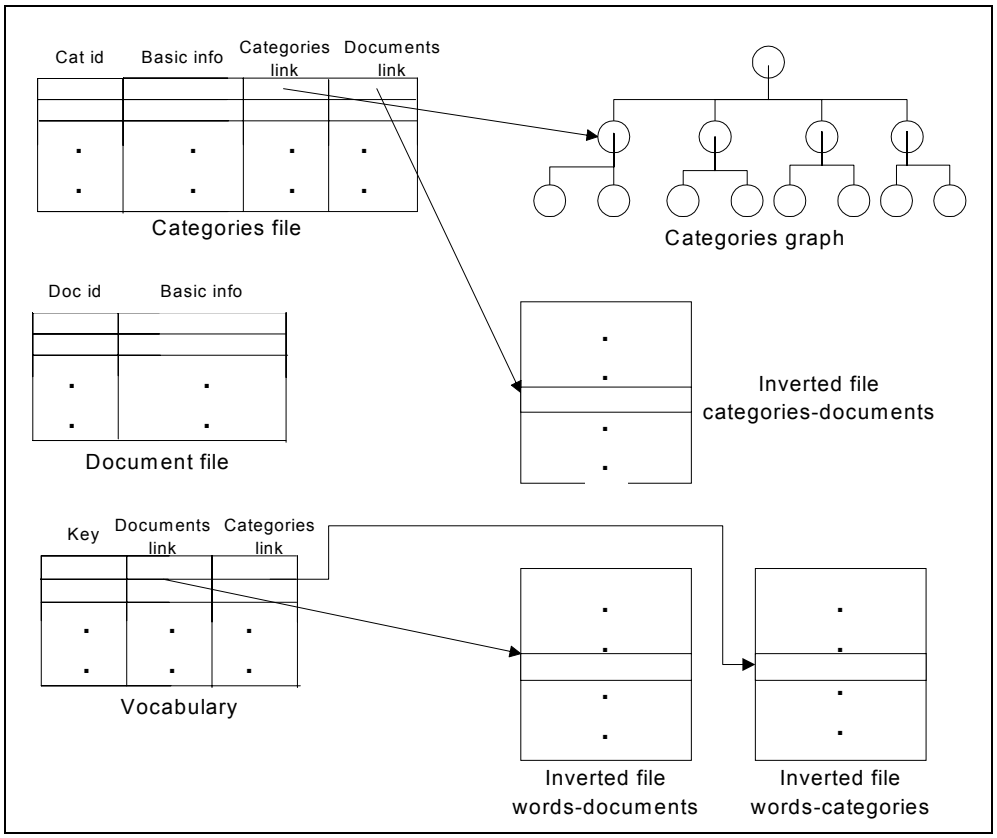

Figure 1: Data structure of the basic model

Figure 1 shows in a schematic way the data structures used in a basic model of Web directory. This architecture is based on similar models, such as those described in [3], [8], [9] and [10].

2.1 Restricted searches

The data model explained permits to solve both the normal search process and the navigation process in an efficient way using an inverted file structure.

On the contrary, the restricted search process in a category graph area requires a more elaborated access to that information. On the one hand, a standard search is carried out retrieving the results, but the key step consists of determining which resulting documents belonged to the specified graph area. Two alternatives are defined for the filtering process.

The first alternative consists of obtaining the list of documents associated to the specified graph area, which is combined with the list of results in order to obtain the final results. The second alternative consists of obtaining the category list from the restriction area (an easier process than obtaining the document list), and checking the results list sequentially, that is, which documents are located at the nodes of the category list.

However, none of the two solutions solves efficiently the searches which obtain a great number of results and which have been restricted to a wide graph area.

3 HYBRID ARCHITECTURE

The problem of restricted searches lies in obtaining an inverted list of results which undergoes a filtering process based on the value of an attribute (associated categories), based on a complex hierarchical structure.

A model of data structures is proposed based on the second alternative, using the dynamic signature files to filter most of the non-relevant results, and thus, the exact filtering only examines the rest of documents.

In the proposed model each document must have a signature which represents each and every one of the categories it belongs to, directly or indirectly. Signature files will be incorporated to inverted files, creating a hybrid scheme of inverted and signature file.

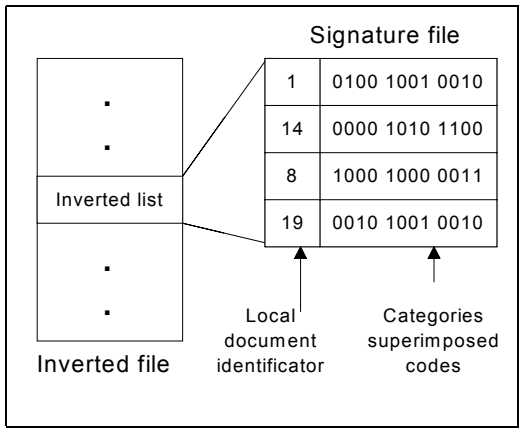

Figure 2: Composed document identifiers in the hybrid data structure

The inclusion of signature files in the inverted lists is due to the possibility of dynamically generating the signature files associated to each query. Thus, when combination operations of inverted lists are carried out the associated signature file will be automatically obtained. Therefore, a composed document identifier has been defined, built with the superimposed signature of every category to which it is associated and with a local document identifier (see Figure 2).

The signature file technique applied to the directed acyclic graph of a Web directory consists of associating a different signature to each category (each node in the graph), and each document shall generate its signature by means of superimposing the category signatures to which it is directly or indirectly associated. The use of superimposing codes for representing hierarchical information is described in detail at [5].

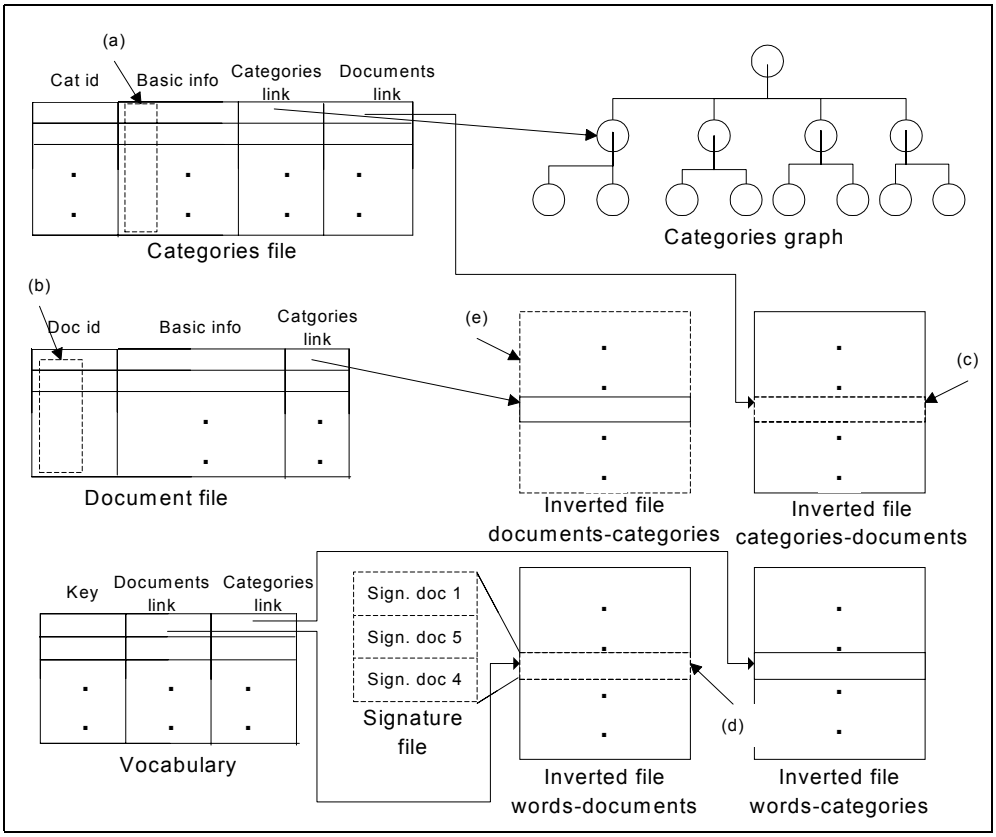

Figure 3: Data structure of the hybrid architecture

4 IMPLEMENTATION

Two variants based on that architecture have been established. The implementations which have been carried out consist of developing a Web directory prototype based on a real environment, integrated by approximately 900 categories distributed into 7 depth levels, in which more than 51,000 Web directories have been classified. The development environment has been an Ultra Enterprise 250 machine with a processor at 300 MHz, 768 MB memory and 18 GB storing space.

4.1 Hybrid model with total information

The hybrid model with total information corresponds to the direct application of the superimposing codes technique to the category graph. In this case, each and every one of the categories have an associated signature.

Figure 3 shows with a dotted line the new data structures or those whose size has increased. As a matter of fact, the increase in size is due to the signatures associated to the categories, specially to the new format of the document identifiers. From a global perspective, the main repercussion of the size increase required lies in the index of key words and documents, while the impact is smaller for the rest of cases. This system supports more than 22000 categories and more than 90 millions of documents.

4.2 Hybrid model with partial information

This variant of the proposed hybrid architecture aims at reducing the size of the signatures and document identifiers, and, therefore, at reducing the storing space. In this case, the number of superimposing operations is reduced by means of applying the technique of superimposing codes only at certain levels in the graph, allowing the rest of nodes to inherit the signatures and the genetic codes of the upper levels.

With regard to the increase in the required storing space, this is significantly smaller, with an approximate reduction of 50% in the required space. It should be noted that the index of key words and documents requires a 65% less storing space. This variant support more than 80 millions of documents and there is no limitation for the number of low level categories.

4.3 Performance evaluation

Five work load situations have been considered for evaluation: void, low, medium, high and saturation. Firstly, the similar performance for the normal searches have been proved.

On the other side, the restricted searches have also been evaluated for the five load levels. Very similar results were obtained for the void, low and medium loads (Figure 4, 5 and 6). The hybrid models clearly improve performance in approximately 50% (retrieving more than 500 results). Oppositely, under a high load situation, the behaviour of hybrid models varies considerably as may be seen in Figure 7. In this case, the performance of the variant with total information worsens significantly whereas the partial information model keeps the 50% of improvement.

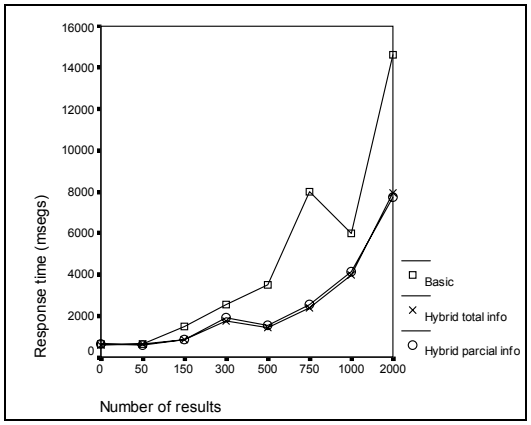

Figure 4: Response time (void load).

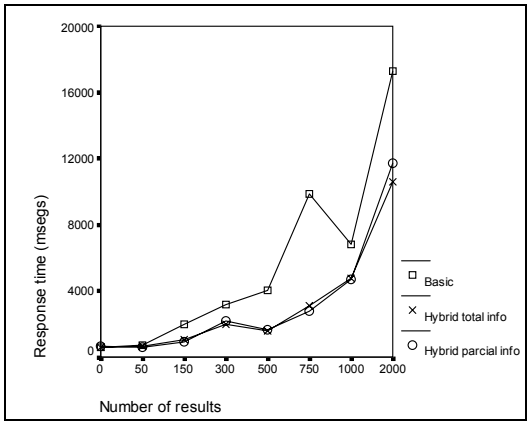

Figure 5: Response time (low load).

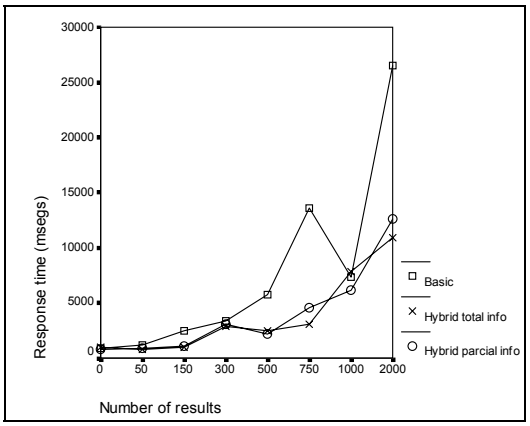

Figure 6: Response time (medium load).

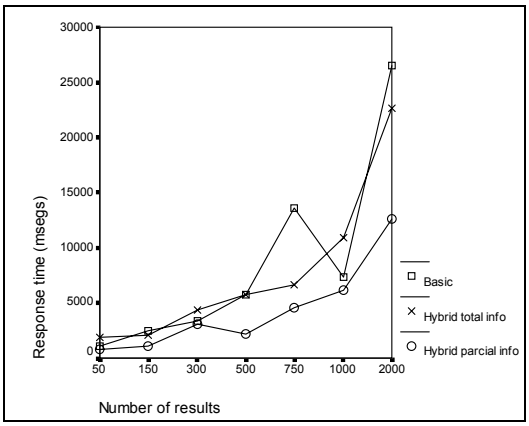

Figure 7: Response time (high load).

5 CONCLUSIONS

This paper describes a hybrid data architecture composed of an inverted file and signature files, especially designed to improve the performance with searches restricted to one area in the category graph.

Two variants of the architecture are defined: the hybrid model with total information and the hybrid model with partial information. The second variant improves performance up to 50% with respect to the basic model under very load situation of the system.

Oppositely, the hybrid model with total information, due to the bigger storing space required, suffers a dramatic decline of performance under high load situations, similarly to the basic model.

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

6 REFERENCES

[1] R. Baeza-Yates, B. Ribeiro-Neto, "Searching the Web". In R. Baeza-Yates, B. Ribeiro-Neto, "Modern Information Retrieval", chapter 13, pps: 367-395. Addison Wesley.

[2] J. Zobel, A. Moffat, K. Ramamohanarao, "Inverted files versus signature files for text indexing". Transactions on Database Systems, 23(4), December 1998, pp.453-490.

[3] S. Brin, L. Page, "The anatomy of a large-scale hypertextual web search engine". The 7th International World Wide Web Conference, Abril 1998.

[4] G. Navarro, "Indexing and Searching". In R. Baeza-Yates, B. Ribeiro-Neto, "Modern Information Retrieval", chapter 8, pp 191-228. Addison Wesley.

[5] F. Cacheda, A. Viña, "Superimposing Codes Representing Hierarchical Information in Web directories". 3rd International Workshop on Web Information and Data management (WIDM'01), in ACM CIKM 2001.

[6] C. Faloutsos, S. Christodoulakis, "Description and performance analysis of signature file methods". ACM TOOIS, 5 (3), 237-257.

[7] S. Stiassny, "Mathematical analysis of various superimposed coding methods". American Documentation, 11 (2), 155-169.

[8] Y. Labrou, T. Finin, "*Yahoo! as an ontology – Using Yahoo! categories to describe documents*". Eighth International Conference on Information Knowledge Management, pp. 180-187, 1999.

[9] W. Lam, M. Ruiz, P. Srinivasan, "Automatic Text Categorization and Its Application to Text Retrieval". IEEE Transactions on Knowledge and Data Engineering, Volume 116, pp. 865-879, 1999.

[10] G. Jacobson, B. Krishnamurthy, D. Srivastava, D. Suciu, "Focusing Search in Hierarchical Structures with Directory Sets". Seventh International Conference on Information and Knowledge Management (CIKM), 1998.