# Computational model for the processing of documents and support to the decision making in systems of information retrieval

**Ph.D Carlos Ortega Maldonado[1], Ing. Paúl Rodríguez Leyva[2], Ph.D Juan Pedro Febles[3], MSc. Hubert Viltres Sala[4], Ing. Yennifer Delgado Mesa[5]**

*[1]Canciller Universidad Especialidades de Espíritu Santo (UEES)*
*Guayaquil, Ecuador*

*[2]Departamento de Soluciones Informáticas para Internet,*
*Universidad de las Ciencias Informáticas, La Habana, Cuba*

*[3]Departamento Metodológico de Postgrado,*
*Universidad de las Ciencias Informáticas, La Habana, Cuba*

*[4]Departamento de Preparación Profesional,*
*Universidad de las Ciencias Informáticas, La Habana, Cuba*

*[5]Centro de Soluciones de Software Libre*
*Universidad de las Ciencias Informáticas, La Habana, Cuba*

---***---

**Abstract -** *Disposing or not, of the necessary information at the right time, can mean the success or failure of any operation. . The field of information retrieval since its inception in the year 1950, has provided tools that allow users to find answers to their needs and questions. In-formation retrieval systems are the most used internationally, since they have interfaces and functionalities easy to understand. The main function of these systems is track the web, store the information found and then respond to user queries. Due to the large amount of information that have search engines, are a rich source of knowledge and support decision-making on information published on the web. Companies like Google do not provide concrete information of which models they use to develop the components of their search engines. In addition the calculation of the relevance of their documents responds to commercial and governmental policies, reason why it is difficult to develop systems as complex as the search engines without owning a computational model that supports the process of development of the same. The present article gives the design of a computational model for document processing and support decision-making in information retrieval systems used to design, development and deployment of searchers at national and international level.*

***Key Words***: *information retrieval, search engine, relevance, computational model*

## 1. INTRODUCTION

Information retrieval (RI) deals with finding material (usually documents) of an unstructured nature (usually text) found in large collections (usually digitally stored) and satisfies a need for information [3]. Information retrieval systems, hereinafter SRI, are tools structured by components that execute functions oriented to the tracing of information accessible on the web, their indexing and subsequently provide visualization mechanisms with which the user interacts directly through queries Which are answered by the system with a list of ordered results according to the criteria implemented in the tool itself.

These systems base their operation on information retrieval models designed to establish mechanisms to respond to users' search needs. The term model has a well-known meaning within the field of information retrieval. This meaning is nothing more than a direct adaptation of its general meaning (theoretical scheme of a system or complex reality that is elaborated to facilitate its understanding and the study of its behavior) to this particular domain. In the field of information retrieval, hereinafter IR, in a simplified way we can consider a model as a method to represent both documents and queries in information retrieval systems and compare the similarity of these representations. To do this, recovery models have to implicitly or explicitly provide a definition of relevance. In addition, they can describe the computational process, the human process and the variables involved in the overall process [3].Companies like Google do

not provide concrete information on what models they use to develop the components of their search engines. In addition, the calculation of the relevance of its documents responds to commercial policies and Governmental organizations, so it is difficult to develop systems as complex as search engines without having a computer model to support the development process. The objective of this research is the foundation of a computational model for the processing of documents and decision support in information retrieval systems that serves for the design, development and deployment of search engines at national and international level.

## 2. PROPOSAL

The proposed computational model is based mainly on 4 principles: scalability, interoperability, standardization and updating. Its structure is composed of 5 components, **figure 1.**

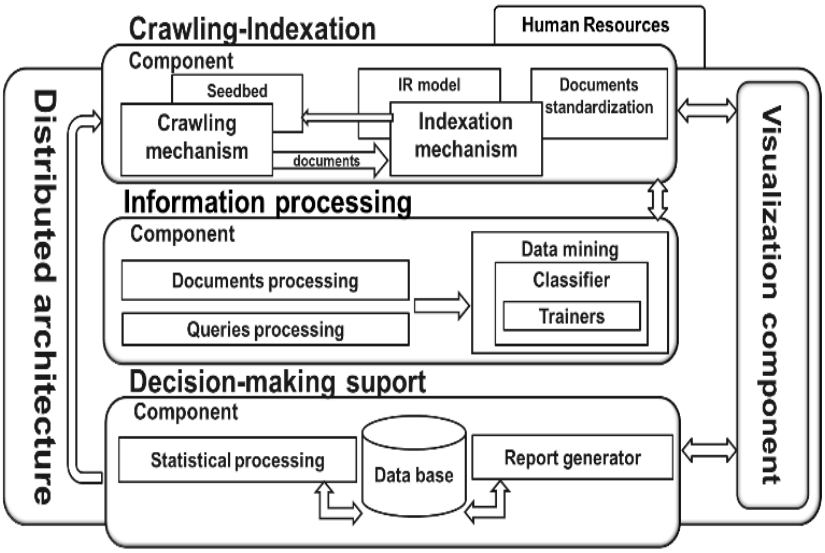


Fig-1: Computational Model. (Own elaboration)

### 2.1 Tracking and indexing component

Periodic crawling and indexing of the web enables SRI to become a source of constant storage. The data stored depends specifically on the internal operation of the search engine and the actual structure of the content hosted on the web. Generally of each indexed document are stored metadata such as: url, content summary, outbound links, keywords, the language in which the document is structured and mime type. The main objective of this component is to track the web and store all the information found; To achieve this is based on two fundamental mechanisms: tracking mechanism and the indexation mechanism.

Tracking mechanism: it bases its operation on a set of web crawlers that initiate a collection process, starting from a seed of links of websites that must contain the greatest possible number of links to achieve a thorough crawl. For its correct operation a distributed architecture is designed, figure 2, where it is proposed to organize the servers taking into account the categories of information of the web to be traced (one server per category) and its size, assessing the processing capacity and Storage of the hardware that is owned. If the web is very large, a small number of crawlers could not meet the proposed goal. The frequency of the tracking process depends on the degree of update of the sites defined in the seedlings of each of the trackers; so users would have the information available in a short period of time after its publication on the web

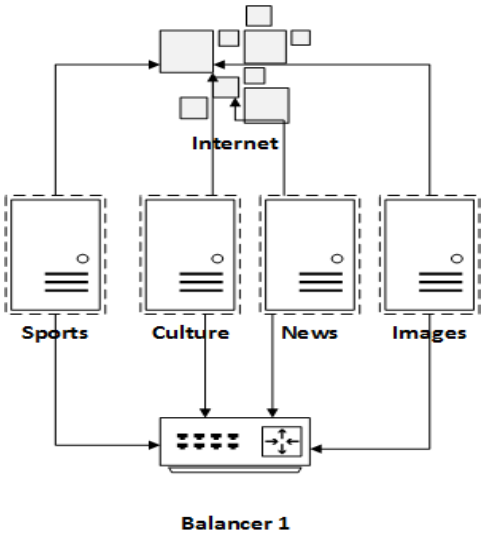


**Fig-2**: Crawler server architecture. (Own elaboration)

Indexing mechanism: is mainly responsible for the structuring and storage of documents. It also applies the information retrieval model for the calculation of the relevance of the documents related to the users' search needs. For its deployment a distributed architecture is proposed, figure 3.

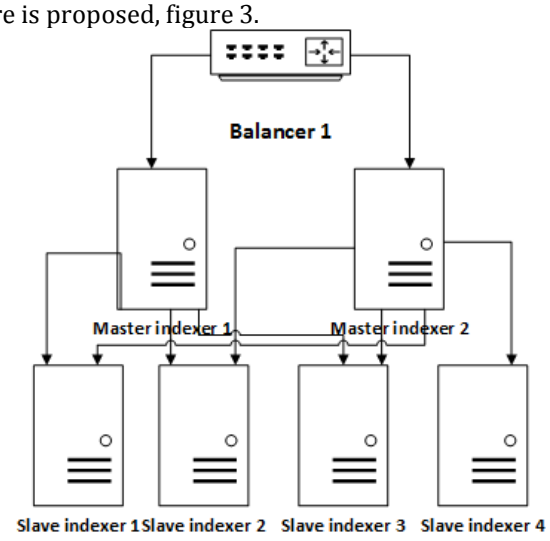


**Fig-3**: Indexing server architecture. (Own elaboration)

Before proceeding to the storage of the documents traced a standardization process is applied that allows to define a correct structure for each one of the documents and avoids to store documents that do not contribute useful information, for which the following metadata is defined as essential:

Mime type: document format (jpg, doc, html, among others)

Summary: a summary of the content of the resource

Date: the date of preparation of the registration

URL: Refers to the electronic address of origin to which the material is attached.

Title: refers to the title that is named by the document

Site: the host from which the document originates

Content: full text of the document

Outbound links: links to external sites

Language: document base language

If the documents to be indexed lack any of the metadata defined above are not stored, to avoid providing incomplete information to the users of the system. The vector information retrieval model is proposed. Salton was the first to propose SRI based on vector space structures in the late 1960s within the framework of the SMART project. Since documents can be represented as vectors of terms, documents can be placed in a vector space of m dimensions, with as many dimensions as components have vector [4]. The fundamental idea on which the vector model is based is to consider that both the key terms with respect to a document and the queries can be represented through a vector in a space of high dimensionality. Therefore, to evaluate the similarity between a document and a query; simply make a comparison of the vectors that represent them [5].

For the computation of the relevance of the documents with respect to the queries, an algorithm, figure 4, is designed to add a variable (SC) to the classical equations of calculation of similarity (Coseno, Jaccard and Dice) that relates the preferences of User searches or search profile (PBU) with categories of stored documents (CDoc). In order to obtain the PBU and thus to establish which category (s) are the most sought by the users and also to obtain the categories of the documents stored CDoc, it is proposed to use web mining techniques as a basis for the operation of the information processing component. Web mining (MW) essentially refers to the discovery and analysis of information from users on the web, with the aim of discovering patterns of behavior [6].

Defined each of the categories on which the categorization process will be based, one should proceed to assign a numerical value to each of them. After the user search preferences are defined as a result of the categorization of each of the previously entered queries, these are sorted according to the predominance percentage (P) of the most consulted categories. If there is only one predominant category, the PBU would be equal to the numerical value of this category. In case more than one category is predominant and have the same value of P, the PBU would be a list with the values of these categories. With the PBU defined and the documents categorized, we proceed to compare the numerical values of the predominant category (s) with the category of each of the documents. If these values match, the variable SC acquires the value of P, guaranteeing that the value of SC is greater when the document belongs to the same group of the category (s) defined in its PBU; otherwise the value of SC is 0 meaning that the category of the document is not related to the PBU [1].

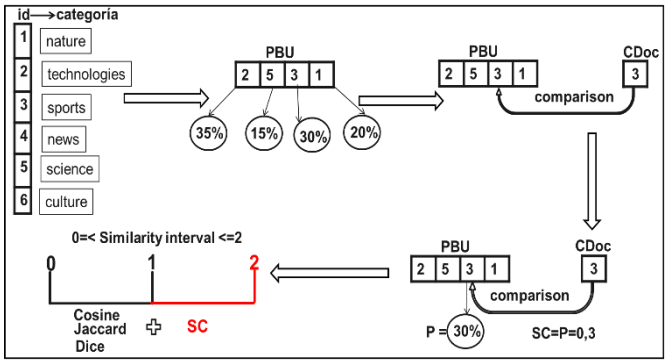


**Fig-4:** Algorithm for the calculation of relevance. (Own elaboration)

Once the value of SC is calculated, this variable is added to the result of the equation of similarity (Cosine, Jaccard or Dice) used in the vector model applied, the resulting value would be the relevance of the document with respect to the query entered by the user. The threshold of similarity initially calculated ranges from 0 to 1, the most relevant documents being the closest to 1. When the variable SC is added and its value added to the initial similarity, the threshold of similarity increases from 0 to 2, more relevant those close to 2. In this way it is guaranteed to provide users with more accurate results and better related to their search preferences [1].

### 2.2 Information processing component

This component performs all the processing necessary for the categorization of each of the indexed documents and the allocation of the search profile to each user of the information retrieval system. A distributed architecture is proposed for its deployment, figure 5. The processing in this component is in charge of two fundamental activities: Document categorization: it is responsible for assigning a category to each of the indexed documents using web mining techniques, which is added as one more parameter in the structure of the document.

Definition of the user profile: it is responsible for classifying the queries inserted by users using text mining techniques and assign to each one a parameter that collects all categories queried by the user.

### 2.3 Componente de apoyo a la toma de decisiones

Su función principal es el procesamiento estadístico de toda la información almacenada en el sistema de recuperación de información que se debe organizar en dos tipos, los documentos almacenados y los perfiles de búsqueda de los usuarios.

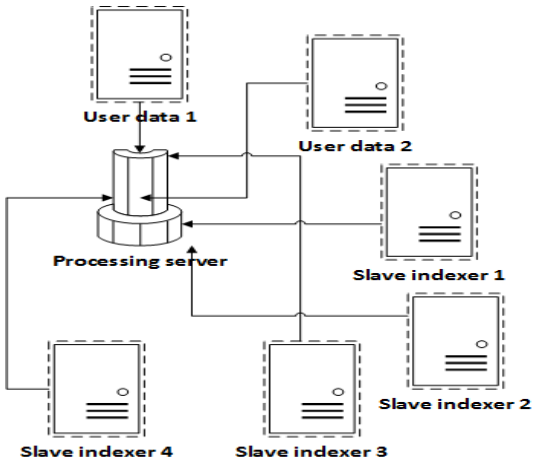


**Fig-1:** Server architecture for processing documents and user queries. (Own elaboration)

The processed information becomes a rich source of knowledge to detect trends of search of the users of the network and thus provide content related to their tastes; In addition to providing data on the publication profiles of each of the websites. The end result of this component is a series of statistical reports that handle data such as:

• User search profile

• Site publishing profile

• Most searched queries

• Search preferences by geographic regions or institutional areas

• Percentage of publication of categories of a specific domain

For its deployment, a distributed architecture, Figure 5, is proposed, where it is represented as the processing server extracts the data relative to the documents of the index servers and the data related to the user profiles of the users data servers.

### 2.4 Display component

It is responsible for providing the necessary mechanisms to users to insert their queries and receive the most relevant results through visualization interfaces. These interfaces have features such as: user registration and advanced search; Which allows to perform a specialized search taking into account 9 main filters [2]:
With any of the words: a search that returns results that contain one or some of the words of the search criteria.
With all words: a search that returns results that contain specifically all the words of the criterion.
With the exact phrase: a search that returns results that contain the exact phrase entered in the search criteria.
Without words: A search that returns results that do not contain any words entered in the search criteria.
Site: allows you to search your own site or domain results.
File type: allows you to obtain files filtered by types grouped in pdf, html, tablets and word documents.
Language: allows you to obtain results in English or Spanish.
Last update: allows you to obtain results grouped by update intervals such as:

- Anytime
- Last 24 hours
- Last month
- Last week
- Last year

Terms that appear: allows you to obtain results that contain the search criteria in different areas of the pages where they were found:

- Page title.
- Contents of the page.
- Page URl

The visualization component is also responsible for providing different visualization options for the statistical reports provided by the decision support mechanism. For its deployment a distributed architecture is proposed, figure 5.

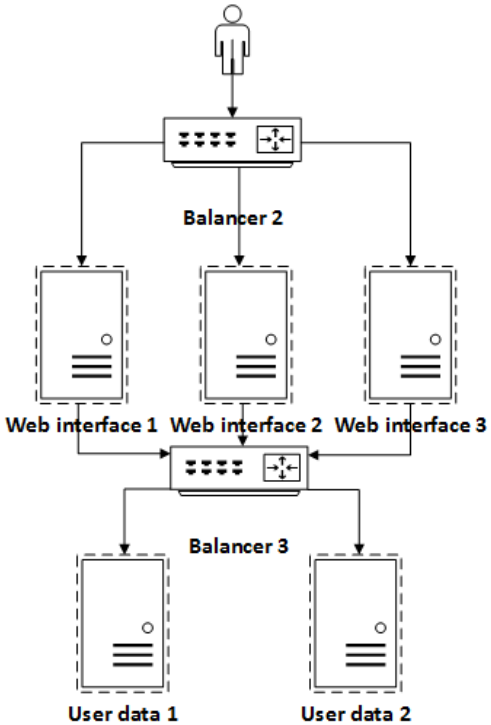


**Fig-5**: Displayed server architecture. (Own elaboration)

The model proposes as indispensable roles for the administration of an SRI and ensure the correct functioning of all the components previously addressed, 5 specialties with the following responsibilities:

Trace Specialist: is in charge of managing the entire process of crawling the web. In addition to maintaining a constant monitoring of the state of each of the tracking servers to avoid any processing overhead and to solve the real-time problems that arise.

Indexing specialist: is in charge of managing the entire process of document indexing ensuring the correct operation of the information retrieval model and checking the document standardization process. In addition, you must maintain a constant monitoring of the growth of the collection of documents in case you need to increase the number of storage servers.

Specialist in information processing: must ensure the correct functioning of categorization mechanisms, checking with small collections of test documents the correct allocation of categories, in addition to a definition of user profiles according to their preferences.

Specialist in information visualization: is responsible for proposing new and current mechanisms to provide information to users. It must process the information provided by the users in the feedback mechanism to propose improvements to the system in an integral way and must ensure the correct performance of the interface display servers and user data storage.

Specialist in statistical processing: in charge of generating the statistical reports requested to support decision making and to monitor the performance of the servers that perform the tasks of this component. The amount of human resources allocated to the administration of the system will depend on the size of the system.

## 3. CONCLUSIONS

The proposed RI model has mechanisms that allow a structuring of documents and user queries that create concrete bases for the process of calculating relevance. The use of web mining to process the information stored in an SRI allows inferring important data such as the search profile of users and the category of each indexed document. An algorithm of relevance calculation was designed that integrates variables such as the PBU and CDoc, which allow users to provide answers that are more related to their search needs. The hardware architecture proposal allows defining an architectural basis for the deployment of SRI at national and international levels. The model is a tool that allows a deep processing of the documents stored in an SRI so it facilitates the process of support to the decision making on the data stored in web search engines.

## REFERENCES

[1] Baquerizo, R., Leyva, P. ,Febles, J. , Viltres, H., Estrada, V. (2017). Algorithm for calculating relevance of documents in information retrieval systems. IRJET, 4,3.
[2] Leyva, P.,Viltres, H., Flores, L.(2016) Componentes y funcionalidades de un sistema de recuperación de la información. Revista Cubana de Ciencias Informáticas, vol. 10, p. 150-162.
[3] N. Diego.(2009) Técnicas de indexación y recuperación de documentos utilizando referencias geográficas y textuales.
[4] S. Mabel. (2013). Buscadores: cómo usar las herramientas de búsqueda en Internet. Informatio. Revista del Instituto de Información de la Facultad de Información y Comunicación, no 2.
[5] Sequera, José Luis Castillo. (2010). Nueva propuesta evolutiva para el agrupamiento de documentos en sistemas de recuperación de información. Tesis Doctoral. Universidad de Alcalá. Página 23, 3 párrafo.
[6] Vásquez, Augusto Cortez; Fernández,Cayo León. (2016). Aprendizaje de perfiles de usuario web para modelizar interfaces adaptativas. Theorema, segunda época, no 3, p. 155-164.

**BIOGRAPHIES**

*Ph.D Carlos Ortega Maldonado Education, Canciller Universidad de Especialidades de Espíritu Santo, Ecuador. Doctor en Jurisprudencia.*

*Ing. Paul Rodríguez Leyva Informático. Jefe de departamento de Soluciones Informáticas para Internet.*

*Ph.D Juan P. Febles Rodríguez Computing Adviser postgraduate, Universidad de las Ciencias Informáticas Cuba*

*Ing. Yenifer Delgado Mesa, Espeialista del Departamento de Soluciones Informáticas para Internet*