International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 03 | Mar -2017 www.irjet.net p-ISSN: 2395-0072

Algorithm for calculating relevance of documents in information retrieval systems

Roberto Passailaigue Baquerizo[1], Paúl Rodríguez Leyva[2], Juan Pedro Febles[3], Hubert Viltres Sala[4], Vivian Estrada Sentí[5]

[1]Canciller Universidad Tecnológica (ECOTEC)
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
[3]Departamento Metodológico de Postgrado,
Universidad de las Ciencias Informáticas,
La Habana, Cuba

Abstract - *This research belongs to the field of information retrieval and its main objective is the basis of an algorithm to assign the value of relevance to a document concerning a consultation inserted by users on information retrieval systems. The concept of relevance is a fundamental aspect in the design and development of information retrieval systems, because although these tools perform a thorough search of the web, a correct structuring of documents and an efficient storage of the same, if the user it does not obtain the results that actually respond to its search needs, then the quality of the information retrieval system is penalized by the acceptance criteria of the users. The algorithm is based primarily on the classical mathematical expressions for calculating similarity between groups, known as the cosine, jaccard and dice formulas. It has the particularity variation of the similarity based on the relationship established between the search profile of users and categories of documents stored in information retrieval system. In order to get these variables are used text mining and web mining techniques allowing the processing of the information generated by the registration of user queries and metadata stored documents? The main contribution of the research is an algorithm to calculate the relevance of the documents that are provided as part of the responses to queries made by users*

***Key Words*: algorithm, similarity, queries, information retrieval systems, relevance**

1. INTRODUCTION

This document is template. We ask that authors follow some simple guidelines. In essence, we ask you to make your paper look exactly like this document. The easiest way to do this is simply to download the template, and replace (copy-paste) the content with your own material. Number the reference items consecutively in square brackets (e.g. [1]). However the authors name can be used along with the reference number in the running text. The order of reference in the running text should match with the list of references at the end of the paper. Information Retrieval (IR) is not a new area, but is being developed since the late fifties. However, it now plays a more important role given the value of the information. It can be argued that having or not having the right information in a timely manner can lead to the success or failure of an operation. Therefore, the importance of information retrieval systems (SRI) can handle - with certain limitations - these situations effectively and efficiently [1]. From 1950 to the present many concepts have addressed this particular issue. According to Baeza Yates, one of the most experienced researchers in this field, the term "deals with representation, storage, organization and access to information elements". This concept is defined by Salton as "a field related to the structure, analysis, organization, storage, search and retrieval of information" [11]. Croft estimates that information retrieval is "the set of tasks by which the user locates and accesses information resources that are relevant to problem resolution." Documentary languages, abstract techniques, description of the

documentary object, etc. These tasks. "[4] On the other hand, Korfhage defined IR as" the location and presentation to a user of information relevant to a need for information expressed as a question "[15].

From the concepts previously provided, it can be inferred that IR is the process of locating and storing information that is then accessed by users depending on the search needs of the same. One of the ways most used by users to access the entire cluster of information hosted on the web are search engines. These tools have a set of components that allow the crawling, indexing and visualization of the collected information as part of the responses offered to the users after they execute a query. They can also be defined as software agents, programs that reside in a computer (host) and have the mission of registering, classifying, indexing and storing documents in the most diverse sites in automated form. In addition, they present the possibility of accessing these databases for consultation. They order the information [14].

A searcher employs different algorithms and methods to satisfy the need for information raised by a user in a natural language query specified through a set of keywords [6, 2]. Algorithms that provide relevant documents in a web search system need to perform an analysis of these documents. This process is based on the generation of an index of relevance between documents for the terms included in them. Generally the analysis performed to generate this index is performed by the search engine but the retrieval of the information is done by a web crawler or web spider [9]. In an interview with Baeza Yates, this researcher explains the following referring to the problematic of the IR, "the current objective is to try to understand the intention after the search. That is, what the person wants to do and personalize their search for that task. "However, classic information retrieval models do not handle variables such as user search preferences; so it is necessary a way to integrate this variable to the calculation of similarity that defines the relevance of the documents in such models.

2. MATERIALS AND METHODS

Irjet Template sample paragraph .Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable. Search engines or information retrieval systems have a number of components, each with a specific function, as shown in Figure 1. Many of these tools base their operation on the classic models of information retrieval. They are recognized as: Boolean Model: The Boolean model is one of the simplest information retrieval models to use. It is based on Boolean set theory and algebra. In this model the queries are logical expressions in which the operands are properties of the documents. Pure Boolean operators are conjunction, disjunction, and negation (AND, OR, NOT). In addition, most systems include proximity operators and simple regular expressions. In general, in these systems the user refines the query until the number of relevant documents retrieved is reasonable [7].

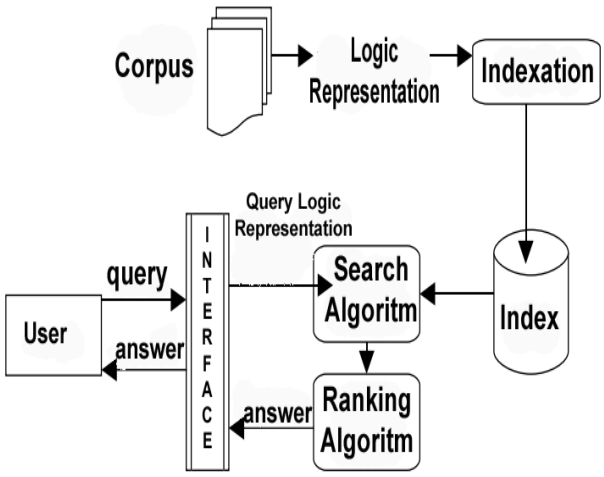

Fig – 1: Basic architecture of a search engine [15].

Probabilistic model [10]: This model is based on the fact that in the IR process it is intrinsically imprecise. Within the process itself, there are certain aspects that are non-deterministic, for example: ⎪ The representation that makes a query of the need for information of the user. ⎪ The representation of the documents in the system. With this in mind, the probabilistic model postulates that the best way to represent this is through probability theory. This model attempts to estimate the probability that, given a query q, a document d is relevant to that query. This is denoted as: P (Rel | d). The model tries to obtain a set of relevant documents (called R), which should maximize the probability of relevance. A document is considered relevant if its probability of being relevant, P (Rel | d), is greater than the probability of not being relevant, P (noRel | d) [3]. Vector model: Salton was the first to propose SRI-based vector space structures in the late 1960s within the framework of the SMART project. Since documents can be represented as term vectors, documents can be placed in a vector space of m dimensions, with as many dimensions as components have the vector [13]. 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 [12]. The three models have algorithms that allow ordering documents taking into account the relevance in relation to a query inserted by the user. In this model the useful words are selected, which are usually all the terms of the texts, except the empty words; this process is enriched using techniques of tagging and labeling. The similarity in the vector model corresponds to the angle between the vector of the document and the vector of the query. If the angle between both vectors is 0 °, they are identical, whereas if the angle is 90 °, they have

absolutely nothing in common. The main advantages of the vector model are the following [10]: ⎪ Allow partial hits, since a document can be considered relevant even if it does not include all the terms of the query. ⎪ The ordering of results is based on several factors: frequency of terms, importance of terms and without prioritizing longer documents. In addition, it allows an efficient implementation for large collections of documents. The documents are located in dimensional vector spaces defined by the same terms. Thus if each term defines a dimension and the frequency of that term defines a linear scale along that dimension, queries and documents can be represented by vectors in the resulting space. Each term of the queries and documents is weighted by assigning them the inverse value of the frequency of the term in the documents of the collection (IDF - Inverse Document Frequency). This value is calculated using the following equation [8]:

$$idf_t = \log(\frac{N}{n_t})$$

Equation 1. Calculating the idf

Where N is the number of documents in the collection and n_t is the number of documents where the term t appears. The weight of a term t in the document vector is given by the equation: [8]

$$W_{t,d} = tf_{t,d} \, X idf_{t,d}$$

Equation 2. Calculation of the weight of a term

Where, $f_{t,d}$ is the absolute frequency of the term t, in the current document.

The search algorithm accepts as input a query expression or query of a user and will verify in the index which documents can satisfy it. Then, a ranking algorithm will determine the relevance of each document and return a list with the answer. It is established that the first item of said list corresponds to the most relevant document regarding the query and so on in decreasing order [15]. With the elements for the retrieval of documents, we can calculate the similarity between a vector of weights of the terms of the query q, and a vector of weights of the terms of document d, with the following equations [8]:

$$sim(q,d) = \frac{\sum W_{t,d} * W_{t,q}}{\sqrt{\sum_t W^2_{t,d}} * \sqrt{\sum_t W^2_{t,q}}}$$

Equation 3. Cosine

$$sim(q,d) = \frac{\sum (W_{t,d} * W_{t,q})}{\sum W^2_{t,d} + \sum W^2_{t,q} - \sum (W_{t,d} * W_{t,q})}$$

Equation 4. Jaccard

$$sim(q,d) = \frac{2 \sum W_{t,d} * W_{t,q}}{\sum W^2_{t,d} + \sum W^2_{t,q}}$$

Equation 5. Dice

With the application of these equations we obtain a similarity value between 0 and 1 that allows assigning a relevance to a document with respect to a query.

3. RESULTS AND DISCUSSION

The algorithm proposed below allows to add to the equations of similarity presented above a variable (SC) that relates the user search preferences or search profile (PBU) to the 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, the use of web mining techniques is proposed. Figure 2. Web mining (MW) refers essentially to the discovery and analysis of information from users on the web, with the aim of discovering patterns of behavior [16].

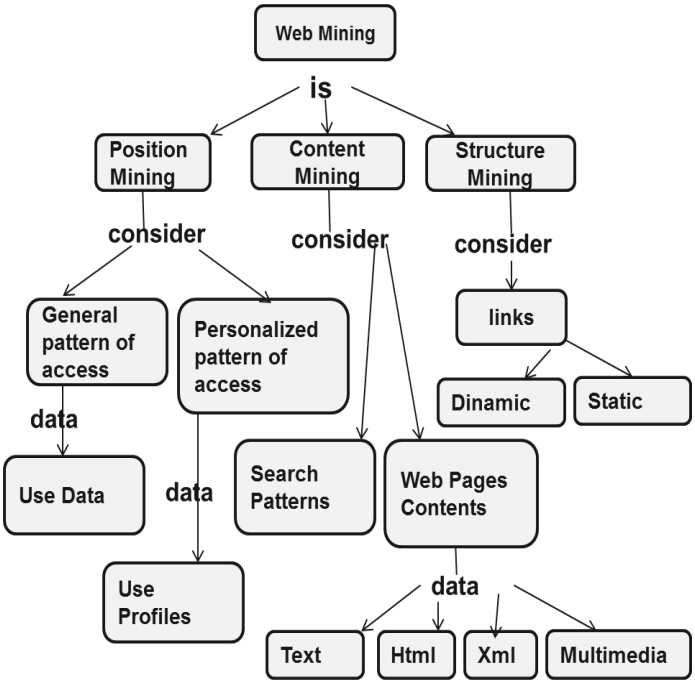

Fig - 2: Web mining process [5]

After defining each of the categories on which the categorization process will be based, proceed to assign a numerical value to each one of them. Figure 3.

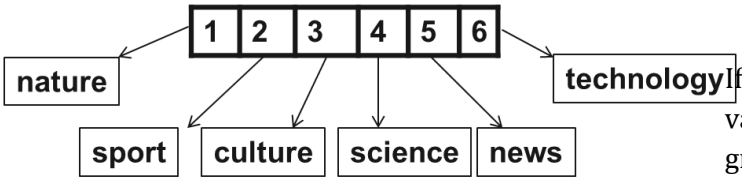

Fig -3: Allocation of numerical value to the categories (Authors´ elaboration)

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 percentage of predominance (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. If 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 Figure 4. With the PBU defined and the documents categorized, we proceed to compare the numerical values the predominant categories with the category of each documents Figure 5.

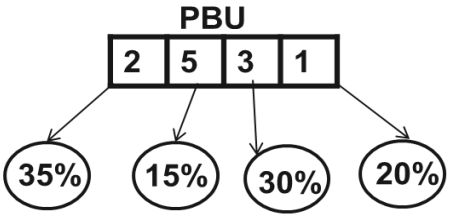

Fig -4: Example of user profile (Authors´ elaboration)

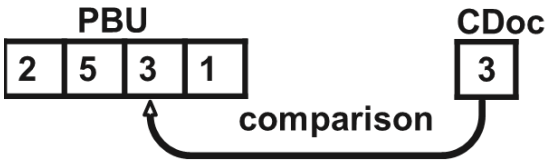

Fig -5: PBU- CDoc comparison (Authors' elaboration)

If these values coincide, the variable SC acquires the value of P, Figure 6, ensuring 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, Figure 7.

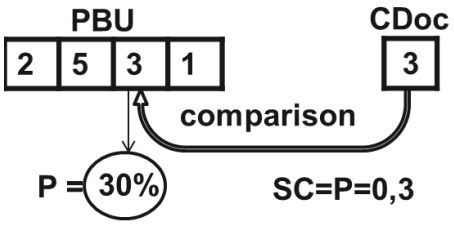

Fig -6: Value assignment to SC (Authors´ elaboration)

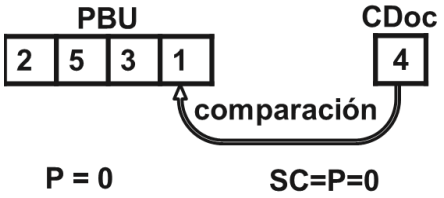

Fig -7: Example of value of SC = 0 (Authors´ elaboration)

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 SC variable is added and its value added to the initial similarity, the threshold of similarity

increases from 0 to 2 Figure 8, the most relevant documents are those close to 2. In this way it is guaranteed to provide users with more accurate and better results related to their search preferences.

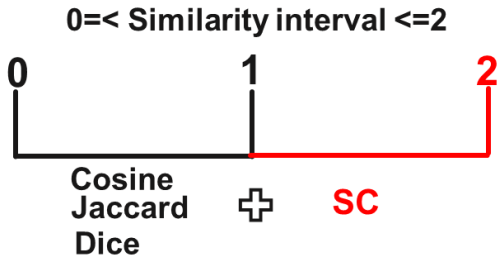

Fig -8: Similarity threshold (Authors´ elaboration)

4. CONCLUSIONS

As fundamental conclusions of the investigation it can be said that:

The vector IR model has mechanisms that allow a structuring of documents and user queries that create concrete bases for the process of calculating relevance. The classic formulas for calculating similarity do not by themselves satisfy the problem of assigning relevance to documents in an SRI. 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. A proposal of relevance calculation was obtained that integrates variables such as the PBU and CDoc, which allow users to provide answers that are more related to their search needs.

REFERENCES

[1] Baeza-Yates, r. Y Ribeiro-Neto, b. (1999). *Modern Information Retrieval. ACM Press.Addison Wesley.*

[2] Blázquez Ochando. (2013). Manual Técnicas avanzadas de recuperación de información: procesos, técnicas y métodos. Madrid. ISBN 978-84-695-8030-1.

[3] Cacheda, M. (2008). *Introduction to the Classic Models of Information Retrieval*. Revista General de Información y Documentación. (18):365-374.

[4] Croft, W.B. (1987). *Approaches to intelligent information retrieval. Information Proccesing & Management*, 23, 4, pp. 249-254.

[5] Fuentes Reyes, Sady C.; Ruiz Lobaina, Marina. (2007). Minería Web: un recurso insoslayable para el profesional de la información. Acimed, vol. 16, no 4, p. 0-0.

[6] Jaimes, l. G y Vega Riveros. F. (2005). Modelos clásicos de recuperación de la información.

[7] Machado García, Neili. (2015). Uso de la similitud semántica para la recuperación de información geoespacial. Página 18.

[8] Monsalve, Luz Stella Garcia. (2012). Experimento de recuperación de información usando las medidas de similitud Cosine, jaccard y Dice. TECCIENCIA, vol. 6, no 12, p. 14-24.

[9] Pino Toledano, David. (2014). Creación de un *crawler* semántico y distribuible para su aplicación en un buscador web.

[10] Reyna, Yunior César Fonseca. (2012). Recuperación de la información: taxonomía de sus modelos. Revista Cubana de Ciencias Informáticas, vol. 6, no 2.

[11] Salton, G. Y MC Gill, m.j. (1983). *Introduction to Modern Information Retrieval. New York. Mc Graw-Hill Computer Series.* Seco Naveiras, Diego. (2009). Técnicas de indexación y recuperación de documentos utilizando referencias geográficas y textuales.

[12] 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.

[13] Seroubian, 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.

[14] Tolosa, Gabriel H.; Bordignon, Fernando R. (2008). Introducción a la Recuperación de Información.

[15] Vásquez, Augusto Cortez; Fernández,Cayo León. (2016). Aprendizaje de perfiles de usuario web para modelizar interfaces adaptativas. *Theorema*, segunda época, 2016, no 3, p. 155-164.

BIOGRAPHIES

Ph.D Education, Canciller University ECOTEC, Ecuador

Ph.D Computing, Adviser postgraduate, Habana, Cuba

Ph.D Computing, Adviser postgraduate, Habana, Cuba