UNIVERSAL DECIMAL CLASSIFICATION – THE SOURCE OF NEW MODELS IN INFORMATION RETRIEVAL

Drd.Mihaela VOINICU
Arges County Library

Abstract: *The current article proposes an information retrieval model based on the librarianship tool called the Universal Decimal Classification. We used a database of documents extracted from the actual database of a library, on which we applied two measures of similarity: the Jaccard coefficient and the Euclidean distance. The novelty of the model presented is that we used a specialty tool developed in the librarianship field, which was later converted into a decimal numerical system, in order to build a mathematical model of information retrieval. Among the results obtained, we may notice the fact according to which the presence of keywords in a query is a necessary but not sufficient condition to get a pertinent response. It is important to identify the perspective from which the query was generated, more exactly the user's fields of interest.*

Key words: *information retrieval, document similarity, mathematical model*

1. Motivation

For several years, within the contemporary info-documentary structures, there has been a real concern to build projects of document management and/or information management. To capture and organize documents and information from/about documents are functions and duties of these organizations.

Within the Information Management provided by the library to the user, we must also take into account factors difficult to quantify such as: user's needs and expectations, on the one hand, but also the possible connexions between information which may lead to new meanings, to fields' interaction and can change the initial perspective, opening the user's curiosity for new subjects, possibly unknown to him/her.

The difficulty to relevantly meet the information needs, namely to get the conclusive information in due time and in an appropriate way, has become today a major problem of all the information and documentation systems.

At the same time, we notice that the role and importance of mathematical modeling within the information retrieval is undeniable.

The main mathematical models behind the retrieval information systems are: the Boolean Model, the Vector Space Model, the Probabilistic Model, and the Language Models [1]. In the recent years, in addition to these classic models, there have been added those models based on clustering [2], latent semantic indexing [3] and the Support Vector Machines [4].

The advantages and disadvantages of these models are presented in [5].

2. Document retrieval model based on UDC

UDC-based search is particularly useful to retrieve titles related to various fields of interest. Thus, to represent fields and areas of knowledge, this information coding system has several advantages, such as: it implements a notation by using the numeric system, it allows the field subdivisions and establishes the hierarchical relationships between knowledge, it allows a short representation of knowledge, a UDC code being shorter than a full text description of the meaning of the code; it removes the language barrier; it provides a much easier approach to applications.

The model proposed in this article combines the vector model of information retrieval with the model based on the conceptual clustering.

The bibliographic description of the documents in the database or in the online catalogue of any info-documentary structure contains fields such as: title, author, year of publication, classification index and key word. The appropriate classification index for each document can be recorded in the decimal system. The numerical value thus obtained is separated into two components. The supraunitary component defines the field (the category) of the document. The subunitary component defines the subfield (the subject) of the documents.

To each document we can attach a vector of characteristics, a tridimensional vector:

$$D_i=(x_i,y_i,z_i)$$

where D_i is a document that belongs to the collection of documents C:

$$C=\{D_1, D_2, ... D_n\}, n\in\mathbf{N}, i\leq n$$

The vector D_i is build as it follows:

-*axis Ox* contains the query terms. The values that the variable x_i can take are:

-0 (there are no query terms in the bibliographic description of the document);

-0.5 (there are query terms either only in the title or in the keywords; the equivalent of the relationship OR);

-1 (there are query terms both in the title and in the keywords; the equivalent of the relationship AND).

-*axis Oy* contains the fields of knowledge, as they are described and used in the librarianship. Thus, the document belongs to a field marked from 0 to 9, e.g. $y_i\in\{0,1...9\}$. If $|y_{i+1}-y_i|>1$ it means that the documents belong to different fields.

-*axis Oz* is the axis of topics. It has values (z_i) between 0.00 and 0.99., e.g. $z_i\in\{0.00,...,0.99\}$. The shorter $|z_{i+1}-z_i|$ is, the closer (the more similar) the topics are.

We calculated the similarity of the documents to a document test (variable), by applying the cosine measure, the Jaccard coefficient and the Euclidean distance. In the case of the Jaccard coefficient (d_J), the similarity is maximum for $d_J=1$, whereas for the Euclidean distance (d_E) the similarity is maximum when the distance between two documents is minimum, $d_E=0$.

As for the cosine measure, the results were not conclusive, the measures being very close, within a narrow range [0,1].

The results obtained for the Jaccard coefficient and the Euclidian distance led to visibly distinct values, which were grouped round various fields.

In *Figure 1* and *Figure 2* there are plotted the query results obtained according to the terms: *Internet and Communication.* The query was made in the same database, respectively the same documents. The query terms are kept, but they are expressed from the point of view of two different fields, namely *Computers* for the first query and *Psychology* for the second one.

We can notice how the documents are grouped differently, their similarity depends on both the query terms and the field the user wants to receive his response from.

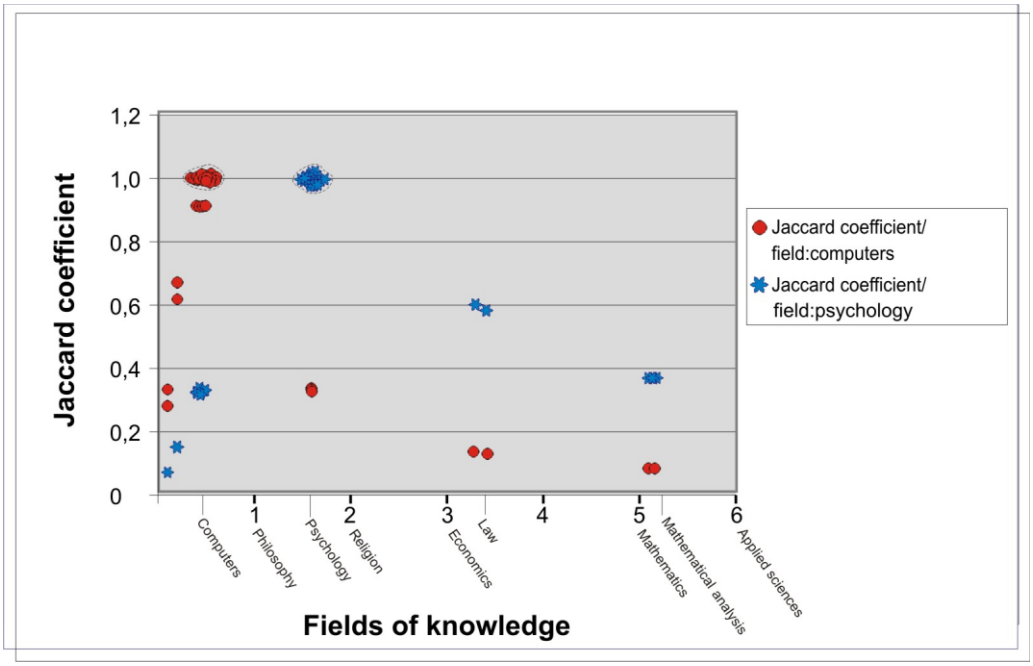

Figure 1. *Document clusters, forming different responses to the same query terms, if the field varies. Relevant response for the fields: computers and psychology (Jaccard Coefficient).*

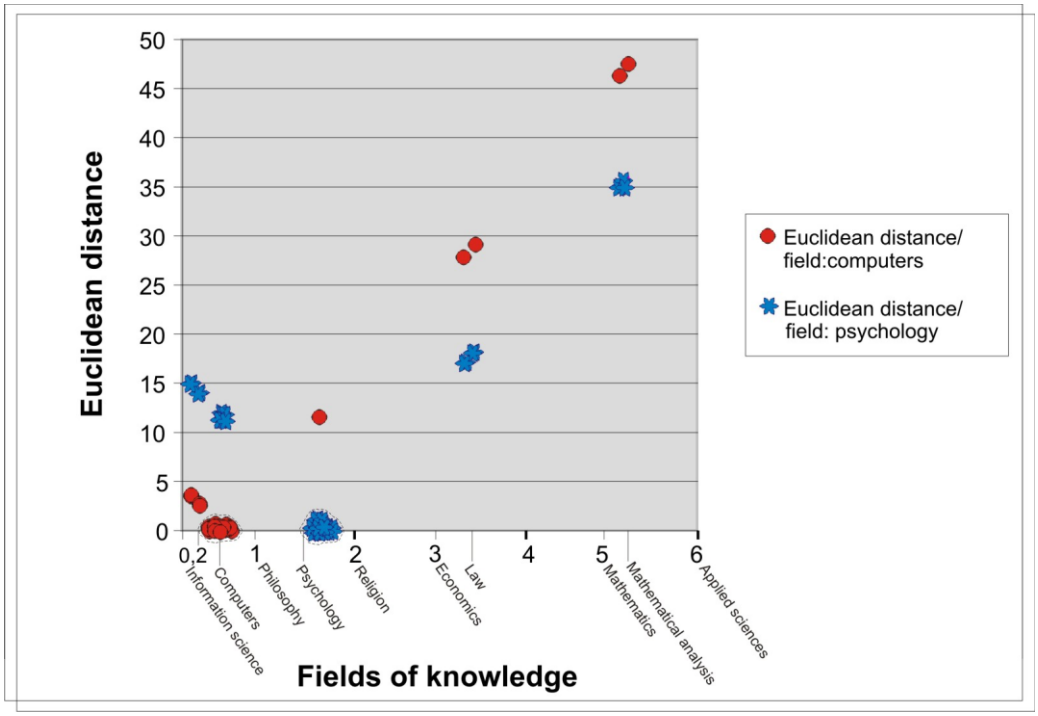

Figure 2. *Document clusters, forming different responses to the same query terms, if the field varies. Relevant response for the fields: computers and psychology (Euclidean Distance).*

3. Results

The bibliographic classifications that exist in the catalogs of the info-documentary structures are important due to the value of their semantic structure for organizing knowledge in the digital global context. The bibliographic classifications are also valuable in terms of the quantity of intellectual work incorporated in the traditional library catalogs, mainly the systematic catalogs, and an issue that should be re-evaluated and reused.

Among the experimental results obtained after using the Universal Decimal Classification to build the current model of information retrieval, we can mention as particularly useful the following:

-obtaining some groups (clusters) of similar documents;

-obtaining some lists of documents similar to one another, but „separated" from the query terms, documents that may be suggestions and recommendations for reading, from different fields, relatively unknown to the user.

The results thus obtained can be included into a software application of bibliographic references, an application useful to any info-documentary structure.

BIBLIOGRAPHY

1. Baeza-Yates R.; Ribeiro-Neto B. *Modern Information Retrieval.* ACM Press/Addison-Wesley England, 1999. pp. 24-38.
2. Gorunescu, Florin, *Data Mining. Concepte, Modele şi Tehnici*. Cluj-Napoca: Ed. Albastră, 2006. pp. 229-244.
3. Landauer, Thomas K.&al. Handbook *of Latent Semantic Analysis*. New Jersey: Lawrence Erlbaum Associates Inc. Publishers, 2007. pp. 3-57.
4. Morariu, Daniel I. *Text mining methods based on Support Vector Machine*. Bucureşti: MatrixRom, 2008. pp. 64-84.
5. Voinicu, M., Jurian M. *Rolul modelelor matematice şi al ontologiilor în managementul informaţiilor*, Braşov: BIBLIO 2010 -Conferinţa Internaţională de Biblioteconomie şi Ştiinţa Informării, 2010. 6 p.