informationresearch

VOL. 17 NO. 4, DECEMBER, 2012

The ecosystem of information retrieval

J.V. Rodriguez Muñoz, F.J. Martínez Méndez and J.A. Pastor Sanchez
Department of Information and Documentation, University of Murcia, Campus of Espinardo, Murcia, Spain

Introduction. *This paper presents an initial proposal for a formal framework that, by studying the metric variables involved in information retrieval, can establish the sequence of events involved and how to perform it.*
Method. *A systematic approach from the equations of Shannon and Weaver to establish the decidability of information retrieval systems and Conrad's equation is used to interweave the ecosystem components.*
Analysis. *This work was developed from a detailed analysis of the scientific literature on information retrieval and through a set of inductive processes; it has been possible to build each of the components of this proposal.*
Results. *First, we have proposed a simple way of assessing the implications of the sequence of events that occur in a search related to the overall response. Secondly, we obtain a formal equation that determines all the interrelated human and technological elements in the information retrieval processes.*
Conclusions. *The establishment of this formal framework allows us to ascertain why we evaluate information retrieval and how one has to intervene in times of imbalance.*

change font

Introduction

If we consider the capacity of the human brain to store information we realize just how ridiculous it is to build artificial devices for the storage that information. This reflection encompasses a certain paradox, since in spite of our brain's huge capacity, it is an extremely complex organ, whose functioning today, at the beginning of the twenty-first century, we are barely beginning to understand. Just as researchers in neuroscience seek to discover the memory and logical functioning mechanisms of the brain, so we in our area strive to optimize and improve the storing and processing of information. There are also, of course, computers (simple machines, however complex they might appear at first sight), that are the paradigm of what is called the complexity paradox, i.e., the simpler it is to use and interact them with them, the more complex their physical and logical architectures are. Their almost childish simplicity at the outset has changed immensely up to the present, not to mention what the future may bring with the development nanotechnology and its advances, which will spur them on to new levels of complexity which today are almost unfathomable (indeed, their builders remind us that it is all but impossible to trace an error within the logic of a microprocessor).

In this scenario the question arises as to what we really have available for processing information. The answer is straightforward: a brain which we only exploit a small percentage of, because we remain ignorant of the structures of its operative system, and a series of machines that can solve a small percentage of our information needs. It is from these perspectives that we must address data storage in computers. It is abundantly clear, too, that the aim of storing information is for it to be retrievable by all who might need it. Databases (regardless of the running model or system employed) seemed to be a first step towards solving the problems of information retrieval; they provide unlimited storage capacity and easy retrieval mechanisms. So, what prevents the problem from being solved completely? The answer is the crucial issue of reductionism: to consider that a part of the surrounding reality can be reduced to a string of simple, semantically unambiguous characters is to run away from that reality.

So the development of more complex systems was undertaken, though even today they remain a long way from solving the problem, in order to move towards a goal that was more in line with the information needs of human beings – information retrieval systems. Since their creation there has been an awareness of a distance between reality and the solution proposed, between the total amount of information stored and the part that is delivered to the user, between the need and the response (Blair 1990). Thus, to a greater or lesser extent, the user will always feel some uncertainty about whether his or her information needs have been responded to as well as possible; so it is hardly surprising that as these systems developed they were accompanied by mechanisms to measure this distance so as to be able to evaluate the retrieval performed by the systems (Martínez Méndez 2002).

The World Wide Web (Berners-Lee, 1989) has become the normal context for the development of these information systems. The dynamics of its functioning and structure have meant that many earlier developments are now sterile on account of the idiosyncrasies of this habitat, including a large part of the measures for evaluating the behaviour of the technology. Researchers the world over are, therefore, in full creative swing as they put forward new ideas and add complexity in an attempt to reduce the distance, to better meet users' needs, ideas that emerge from the mind, with the response of artificial devices.

Manzelli (1992) helps us to understand the arguments and the aspects surrounding retrieval systems and the parameters according to which they must be deployed, as well as the ups and downs to be confronted for them to evolve along lines that are more cohesive with human thinking (Darwin himself stated in 1859 that for any system, however complex, to survive or evolve, it had to transform so as to adapt better to its environment, to its own and to external demands). According to Manzelli:

> *starting from the premise that human understanding can only capture information processed by the evolutionary functions of the brain, we can understand the traditional scientific paradigm is incomplete since it establishes an arbitrary division of the reality of the subject that observes and the object that is under scientific observation. The brain is an objective reality because it is part of nature. (Manzelli 1992)*

Nowadays, the processing of information and, more specifically, all transformations to which we subject human knowledge '*in relation to sources and channels of information, including both active and passive information seeking, and information use*' (Wilson, 2000), are not subjected and constrained to human brain processes alone, rather it is the information technologies that play a decisive role. Manzelli (1992) points out, that, albeit with greater prudence, we can understand that the independence postulated by Descartes between *res extensa* and *res cogitans* is being overcome by the connections that exist between human thinking and the automatic processing of information. Indeed, the Web and, even more so, what is known as the *semantic Web*, has joined this course in its interaction with the search for new meanings in human beings' information needs; it has moved on from the merely readable, the data (the abstraction of the reality through thought) to the intelligible (the interpretation of this reality by the system itself), all of which is achieved by gaining information from the system, in this case by adding code.

Manzelli goes further when he says that just as the brain functions by using matter, the brain cells of the neuron system and energy, bioelectricity, to process information, so technologies process information with chips and electricity:

> *if we wish to incorporate a new advanced scientific paradigm to the brain's functioning (or to other information or negative entropy operators, like chemical catalysts or biological enzymes) as a part of the observed reality, then we need to include Information as a general conception for a complete description of the transformation processes in contemporary science. (Manzelli 1992:)*

With this approach to the need to incorporate information

> *as a new general physical conception of global reality, we can symbolize this new paradigm as $< E|I|M >$, i.e., include the notion of information (I) in the fundamental and general variables that science uses in the general description of the interaction between Energy systems (E) and Matter (M) (Manzelli, 1992).*

This equation that includes the brain and/or any type of natural or artificial extension susceptible to processing information as an observable physical reality should be considered as a global system paradigm, as a conceptualization of a situation in which the meaning of the objective reality is enlarged to include the subject (knowledge, computer and any device for elaborating information) that works to observe and interpret the object's reality, in this case information, or in its interpretive version, data as a global vision of the systems. Thus, Manzelli (1992) furthers our purpose by indicating that,

> *using the paradigm $< E|I|M >$, we go beyond the old mechanical conception of scientific interpretation, and we can return to the old ideas about the intelligence of nature. Indeed, the Greek philosophers did not believe that intelligence was a function of the brain, but a purpose of the of the natural system as a whole; similar to the paradigm $< E|I|M >$ we consider creativity as a function of the whole universe and we see knowledge of the whole universe as a generalized system of learning which develops in the sense of natural intelligence. (Manzelli 1992)*

The same author continues with the following argument: if a fundamental principle establishes that energy is neither created nor destroyed, a logical consequence is that the total energy must remain constant.

$$\mathbf{E}_{\text{TOTAL}} = (\mathbf{E}_{\text{L}}) * (\mathbf{E}_{\text{M}}) + (\mathbf{E}_{\text{I}}) = k \ \mathbf{(1)}$$

If, in this scenario, we see information as a variable of the general description of the transformation of energy and matter, the overall variation global, δ, of the various parameters of energy obtained through equation (1), then the component E_{TOTAL}, which includes the dissipated fraction within the informative processing of the brain, (or of other coding and decoding processes of information) must be equal to zero, since at all times $\boldsymbol{\delta}\ [(E_L) + (E_M) + (E_I)] = 0$ and, hence:

$$+\boldsymbol{\delta}\ (\mathbf{E}_I) = -\delta(\mathbf{E}_L) - \delta\ (\mathbf{E}_M)\mathbf{(2)}$$

Consequently, the positive increase of information can be seen as an evolutionary programme in which nature progressively transforms the interactions of energy and matter to develop a growth of energy gained from processing information, ($+E_I$), i.e., a loss of chaos, or an entropic decrease which Manzelli (1992) calls *Principle of Creative Evolution*, or the tendency of systems to transform matter and energy to gain in degrees of information.

Thus, the processes of interaction described in terms of < E|I|M > can be seen an evolution of the overall degree of quality of the system, which is not only referred to measurable mechanical processes but which is a holistic vision of the system. As with information technologies, an equivalent procedure is observed when the software is incorporated into the computer hardware, with the ensuing progress of the following generations of computers, like in nature, to achieve the progressive increase of information in the energy-matter transformation processes and so gain a good intuitive expression for interpreting the evolution created.

With formula (2) we can explain the main a direction of the evolutionary processes. In fact, if we do not consider the objective existence of information as a physical aspect of nature, it is impossible to refer to a complete meaning of objective interpretation of global events that are produced in our environment. Hence, this line of reasoning allows us to argue that Heisenberg's *uncertainty principle* is a consequence of ignorance of the consideration of the parameter information in explaining physical phenomena; disinformation is an imperative, subjective reality of the observer. The question then is to establish the importance of the phenomena that are to intervene in information retrieval systems, the consequences of its evolution and the necessary evaluation processes for an improvement that increase usefulness, i.e., $+\boldsymbol{\delta}(E_I)$, considering these within the paradigm < E|I|M >. This paper expresses the current situation of information retrieval systems: they are unfinished devices.

Evaluation of information retrieval systems

The determinist nature of information retrieval systems requires them to be evaluated. Hence, alongside the development of their technology there appears a broad field of work devoted specifically to determining measures that value their effectiveness (Martínez Méndez, 2002). An exhaustive review of the specialist literature identifies several evaluation groups: those based on the importance of the documents; user- based ones, and a third group of alternative measures to those made by judgments on relevance, that seek to avoid being affected by the amount of subjectivity inherent to these judgments (Martínez Méndez and Rodríguez Muñoz 2003).

Like any other system, retrieval systems are susceptible to being evaluated so that users can measure their effectiveness and so acquire trust in them. Pors notes that the evaluation of information retrieval systems have '*a long tradition in information retrieval research, especially in the so-called system-oriented tradition*'. (Pors 2000: 59)

For Blair (1990: 69) '*the field of information retrieval has a critical need for testability, just another field that aspires to scientific status does*'. Likewise, Baeza-Yates (1992) states that,

> *an information retrieval system can be evaluated in terms of many criteria, including execution efficiency, storage efficiency, retrieval effectiveness and the features they offer a user. (1992: 10)*

Effectiveness and efficiency are at times confused. Effectiveness in execution refers to the measure of time taken by an information retrieval system to perform an operation (ths parameter has always been the main concern in such a system, especially as many of them have become interactive and a long retrieval time interferes in the usefulness of the system and can even dissuade users). Efficiency of storage is the measure of space the system needs to store data (an appropriate value is the ratio of the index file size and the size of the document archives over the size of the document files, known as the excess space, which should always be of values between 1.5 and 3 in systems based on inverse files).

In any case, the same author states that traditionally much more importance has been conferred on evaluating effectiveness of retrieval, related to the relevance of the documents retrieved. Pors holds that

> *experiments, evaluations and research have a long tradition in information retrieval research, especially research on the exact comparison paradigm, which concentrate on improving the terms of question and the representation of documents in order to facilitate increased exhaustiveness and accuracy of searches. (Pors 2000: 59)*

The same author suggests that when evaluating there is a difference between physical access evaluation and logical or intellectual access evaluation, and considers that evaluations have to focus on the second type. Physical access concerns how the information sought is retrieved and represented to the user. It has to do with the way in which an system finds the information, or provides the user with guidelines as to its location. This access is closely linked to techniques for retrieving and presenting information. Logical access is related to locating the information desired. To illustrate the above, Blair profers the following example:

> *consider a library: discovering where the book with a call sign QA76.A1A84 is in the library is a problem of physical access; discovering wich book in the library will be likely to satisfy a particular information need is a problem of logical access. (Blair*

> *1990: 70).*

The latter has to do with the relevance of the object located with the constituting issue of a specific request for information. Thus, Blair considers that problems of logical access are more important than those of physical access, which are solved once previous problems have been.

These statements, made in the early 1990s, are equally valid twenty years on. Pors notes: *'There are several approaches to the study of information retrieval. One is grounded in the sciences and has experimentation as its basis. Another more modern approach is the user-centred or cognitive viewpoint'.* (Pors 2000: 59). ideas which are equivalent to the physical and logical accesses of Blair.

Another fact to consider is the tremendous competition between the alignment algorithm used by different Web search engines. These are systems that compete to present as many documents as possible to users in the shortest time possible, without considering that the user may prefer to wait longer to get back a set of more relevant documents. This tendency to advance in the development of the physical aspect confirms the fears of Blair (1990), who believes that excessive evaluations are made of aspects related to physical access, when what are really required are more evaluations of the logical access. Following this line of reasoning, what should be evaluated in order to determine with certainty that the information a system provides is valid for its users? For Blair (and many others, including ourselves), there is no doubt that logical access needs to be evaluated by analysing the relevance or non-relevance of the document retrieved '*Most IR experimentation has focused on retrieval effectiveness -usually based on document relevance judgments*' (Baeza-Yates 1992: 10).

In an information retrieval system, the retrieved documents will not be exact responses to the request, it will be a set of documents ordered according to the relevance of the request. What needs to be evaluated is how related with the topic of the question the set of documents forming the response is; this is the retrieval effectiveness.

A new question arises here that, while seemingly trivial at first sight, may markedly affect the result of an evaluation process: how can one respond to a question with certainty as to when a document is relevant. The *Diccionario de la Real Academia Española* defines relevance as '*the quality or condition of being relevant, importance, significance*', and the term relevant is '*important or significant*'. By extension of the previous definitions, we understand that a retrieved document is relevant when its contents contain something of importance or significance in terms of the underlying motive of the question made by the user, i.e., with the expression of the user's information needs. Knowing the meaning of the term is not of great help, since new problems arise when determining exactly when a document can be considered relevant or not. We should not forget that these problems are closely related to the cognitive nature of the process, of which the following are highlighted below:

- A document can be considered relevant or non relevant by two people according to the motives for the need for the information or the degree of knowledge both have of the subject. In an extreme case, the same document might appear relevant or not to the same person at different moments in time (Lancaster 1991).
- It is difficult to determine beforehand when a document is relevant and it is even complicated to specify this clearly and concisely. In fact, '*it is easier to proceed towards the determination of relevance than to explain how the same has been done*' (Blair 1990), who also thinks that,

  > *it is clear that it is a very subjective notion which may be explained in a variety of ways by different inquirers (and even by the same inquierer at different times). We should not to be too surprised either if an inquirer, who claims that he knows perfectly well which documents contain information relevant to his request, cannot readily explain just what he means by relevance (Blair 1990: 71)*

  This does not mean that the concept lacks importance, but that judging the relevance is part of a wide set of daily tasks that we carry out in a habitual almost routine manner - cognitive processes, therefore, and contributors of the paradigm < E|I|M >- but ones that we then have problems in finding the right words to describe them.
- Finally, it may be somewhat risky to clarify a document absolutely as being relevant or non-relevant to a subject. It is usual to find documents that are relevant in some of their sections to a certain subject, but not in all. Some authors have taken up the idea of *partial relevance*, as, for them, relevance cannot be measured in binary terms, but can acquire many intermediate values: highly relevant, relevant, barely relevant, minimally relevant, etc., meaning that relevance can be measured in terms of a continuous function rather than a binary one.

These impediments somewhat condition the viability of relevance as an evaluation criterion of information retrieval. Cooper (1973: 88) introduces the idea of the '*utility of a document*', and considers it better to define relevance in terms of the perception a user has of a retrieved document, i.e., whether it is going to be useful to him or not. This new angle has a big advantage: it places the estimation of the suitability or non-suitability of a retrieved document in the judgment made by the user, to the extent that, as mentioned above, by enumerating the problems of relevance, we can assume that a user will have problems in defining what is and is not relevant, but he will have few problems in deciding whether the document seems useful to him or not. It is the user who is going to analyse the document and who is going to use it (if convenient), and so the judgments on relevance will be made by the user, and it is those judgments that are going to decide whether a system is considered good or bad. The importance of this concept leads Blair to conclude that

> *it simplifies the goal of an information retrieval system and though its evaluation is subjective, it is measurable and more faithful to the conduct of document retrieval in ordinary circumstances.(Blair 1990:73)*

Frants (1997) proposes another meaning of relevance, albeit very similar, in terms of *functional efficiency*. A system scores highly here when the majority of the documents retrieved satisfy the information demands of the user, i.e., the user finds them useful. Lancaster (1991) thinks that although another terminology may be used, the word relevance seems the most appropriate to indicate the relation between a document and a request for information made by a user, although it may be wrong to assume that the degree of relation is fixed and invariable; rather, it is better to say that a document has been judged as relevant to a specific request for information. This author reflects along very similar lines to Blair, considering that the relevance of a document with respect to the need for information put forward by a user does not have to coincide with the value judgments made by many experts on the content of the document, but according to the satisfaction of the user and the usefulness that the contents will have for him. Lancaster (1991), finally thinks that the word *pertinence* is a better choice. Thus, relevance will be associated with the concept of the relation existing between the contents of a document and a specific topic and pertinence will be limited to the relation of the usefulness existing between the document retrieved and the individual's information need

The pertinent set of retrieved documents can be conceived as a subset of the documents stored in the system that is appropriate to the information needs of the user. The *Diccionario de la Real Academia Espa&ntile;ola* defines pertinence as the *quality of being pertinent*, understood as pertinent to *all that has to do with it or is opportune*. We can state that a pertinent document is one which is opportune, because it gives the user the information that fulfills a purpose for him. Similar opinions are found in Foskett, who

> *distinguishes between relevance to a request that he calls 'relevance,' and relevance to an information need, that he calls 'pertinence'. The former is seen as a public, social' notion, that has to be established by a general consensus in the field, the latter as a private notion, depending solely on the user and his information need. (Mizzaro 1997: 816)*

Really,

> *'the different approaches have the same ultimate objectives because the evaluation procedures are concerned with the question about*

> *the system's ability to satisfy information needs and to improve the totality of the information retrieval process' (Pors 2000: 61).*

An interesting additional bibliography on this meaning of the concept of relevance is provided by Lancaster (1991) who gathers quotes from Cooper, Goffman, Wilson, Bezer and O'Connor; another compilation has been made by Mizzaro (1998), who cites, among others, Vickery, Rees and Schultz, Cuadra and Katter, Saracevic and Schamber. This set of opinions has been accepted by later authors in this field. Indeed, a special issue of the journal *Informing Science* includes the following sentence by Greisdorf (2000): '*in the last thirty years no practical substitute has been found for the concept of relevance as a criterion for measuring and quantifying the effectiveness of information retrieval systems*'.

In the same vein, Gordon and Pathak (1999) opine that judgments of relevance made by experts are,

> *conforming to accepted IR measurements (line recall-precision curves) to allow results to be evaluated in a familiar context. (Gordon and Pathak 1999:147).*

In reality, the retrieved documents are not relevant or non relevant strictly speaking, in that it is not a binary decision because the contents of the documents can coincide to a greater or lesser extent with the information needs. What can be determined with greater security is whether they are relevant or not for a particular person. Pragmatically, the same document can signify different things for different people; judgments of relevance can only make semantic or even syntactic evaluations of documents or questions. However, these judgments may fail in their generalization because individual users participate in them and they may have problems in identifying where the user really finds a particularly relevant document. These authors take a slightly tongue-in-cheek approach in their paraphrased: *relevance resides in the details*. We therefore assume that the initial approach that a document will be relevant to our needs when it really contributes some content that is related to our request, so when we speak of relevance we will be speaking of pertinence, provided that we are referring to the point of view of the final user performing an information retrieval operation.

Ecoystem

All the reflections and arguments so far have been made in relation to information, by projecting one stage of its lifecycle, retrieval, with the aim to establish a framework of behaviour in which we can formally establish a systemic structure. Retrieval and the need to evaluate this process should, therefore, be placed in a formal context that can establish the limits and parameters in which they unfold both as a system (intrinsically) and with the environment (extrinsically). While some new terms will be presented here as a consequence of our opening up the way with formalisms whose axioms and definitions are not definitive, as

Wagensberg (1994: 51) says: '*there are certain temptations for the thinker; for example to invent aterm to account for a mystery and then slowly and surreptitiously elevate it to the status of explanation*'. However, it is no less true that we find ourselves in an environment of which information is an essential part, as Aladro holds:

> *Information creates reality, and reality, in any of its components, however small or insignificant they may be, spreads like an informative wave... [Reality and information feed back into themselves, one does not exist without the other] Thus, mastering the informative universe supposes mastering the reality in which we live (Aladro 2009:10)*

Information can, therefore, be seen as an *infinite source of energy*.

For a suitable framework for the information in our study, we can say that the discourse universe is information as object not as subject. If we consider as object everything that may be material for treatment and study on the part of the subject, our hypothesis lies in the fact that when we speak of the information-object subspace we are really referring to the data space, while when we speak of the information-subject subspace, we are referring to the knowledge space, as can be seen in the illustrations below.

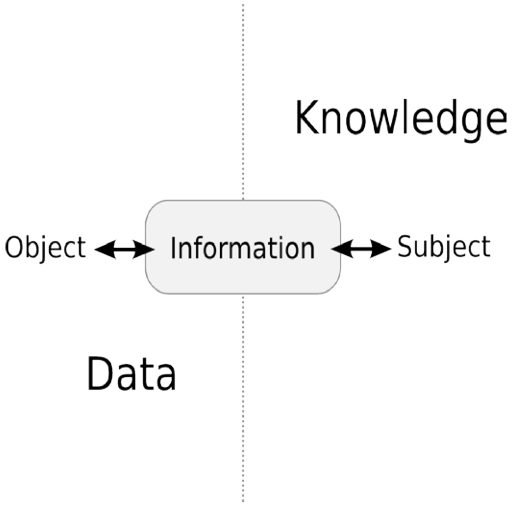

Figure 1: Information as object and subject.

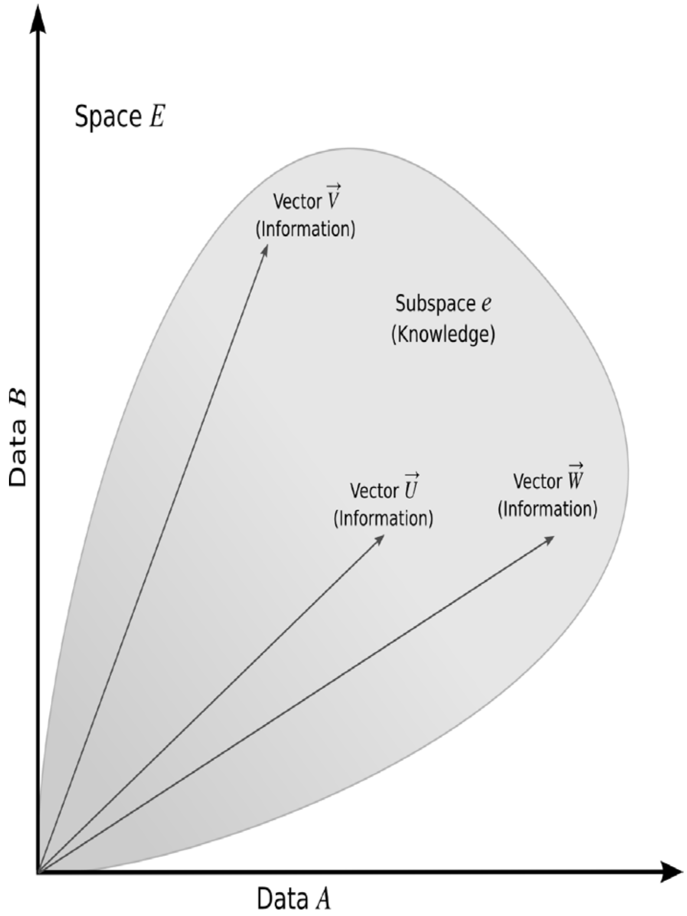

Figure 2: The spatial scenario of information.

Thus, we construct the argument inside a space whose coordinates (orthogonal axes) are represented by the data. We will have an n-dimensional space, where n is the number of data, such that through the composition of these we will obtain a sub-space of vectors that we can identify as information, such that a piece of information is a vector whose components are defined by data. Likewise, we will be able to determine sets of linear dependencies within this vector space, or sub-spaces of sets of linearly dependent vectors to which we can make a particular piece of knowledge correspond.

We now tackle the first question that needs to be solved: the approach. This can be considered atomistically or holistically. Both allow for analyses and reflections of interest. In the first observation, it is worth noting that while the atomistic approach gives a precise description of the elements comprising the system, leading to their definition and, (why not?) classification, it is no less true that it deprives us of a certain height vision, i.e., the rules of synthesis are such that they can change the semantics of the system in question. The holistic approach does allow the system to be observed as a whole. To do this, we will look at the prototype and test its robustness. Coherently, we should note that if we wish to make a deeper analysis we will be able to conclude that we define the object as everything that is susceptible to being described by a subject and, in the same sense, we express the idea of subject as that which is described. In the environment we are working in, our object is the representation, treatment, to obtain new sources of subject-information.

In a universe where physical laws are subject to uncertainty (i.e., by chance), reality is determined by observation, interpretation and informative processes. In the scientific model, observation and gathering of information decree the reality and induce uncertainty to a specific event or state of that reality. Knowing the nature of the information may be determining in knowing how to use it, to make it fulfil a certain end and use it or take advantage of it to dissipate disinformation (Moreiro González 2004). In order to begin this approach we need to enlarge the framework outlined above so as to situate information as the supporting axis of the data and knowledge as agents that are close to and contain the reality.

Ontologically, we will interpret knowledge as the being and data as the not being of any system, where information is the fluid, the essence for transforming uncertainty into reality. It is, in terms that Manzelli (1993) calls *Principle of Creative Evolution*, seeing knowledge as the subject that works to observe and interpret the reality of the object, in this case the interpretative version of information thorough data as a global vision of systems in order to attain knowledge. To understand this we need to adopt a systemic approach in which data are seen as meta-information and information as met-aknowledge in a type of axiom where the vital impulse leans towards an explanation that goes beyond the physical representation proper to be described, and see information as a description of knowledge, and data as a form of information. It is, therefore, to define information as a proto-form of the data, which is lit by knowledge through creativity.

However, we need to go a little further in our analysis of that object. In order to describe reality in technological terms, what in the context of databases is known as the discourse universe, the first characteristic is readability, an essential structural and functional quality of any computational tool. The absence of any syntactic or algebraic ambiguity is indispensable for the objective description of the reality. Since, in this representative scenario we have a static, determinist framework in which all that is observable is perfectly measured, physically we are in a Newtonian context and with a documental perspective that governs the norms of the documental technical process. Within this order of things, we can call this scenario one of information statics or *infostatics*, a closed system in which we can place languages that define data in the databases.

On the other hand, we have the subject, characterized by intelligibility, a dynamic, non-determinist framework caught by the arrow of time, in which uncertainty and semantic ambiguity are dominant in the troubled waters of chance. In this sphere information retrieval systems act, so they are not systems that are measurable directly, but by prediction. From a physical standpoint, we would say that we are in a quantum environment, where the governing laws are subject to undecidability, or where it is not possible to design a system that decides whether the response to an information retrieval operation will be correct. In this case, we can label this scenario information dynamics or *infodynamics*, an open system in which we can place the system, since, for the same representation, the variable contexts change the result.

If the route for the object (data readability) is made and normalized in a digital, technological context, then we must, in this same context, advance towards the subject (knowledge and intelligibility). The dividing barrier is the representative granularity, i.e., Semantics, in the sense of the meaning and interpretation of the data. Today, the attempt to overcome that barrier is done by adding code to the data; one paradigm of this is the semantic Web (Berners-Lee *et al.* 2001); (De Virgilio *et al.* 2010), Pastor Sánchez (2011). Before analysing the ecosystem (the unity composed of interdependent organisms sharing the same habitat), where we may interpret how and what needs there are in adapting information retrieval systems, we have to see if there is a possibility of a measure that indicates the degree of chaos in which information retrieval acts. Unspeakability precludes *a priori* knowledge of a system's response, since in attempting to understand this circumstance, Burgin (2010: 58) states that any statistical measure of information, such as the amount of information (in terms of the Shannon and Weaver equation), entropy of a message **m**, uses probabilistic concepts and constructions, where $\mathbf{p_i}$ is where **pi** is the probability of the event **i** (the outcome of an experiment, case or situation):

$$\mathbf{H}\ [\mathrm{m}] = \mathbf{-}\delta p_i \log p_i$$

In spite of its success and popularity, the Shannon and Weaver statistical theory of information has not been able to solve many problems of information processing. The same occurs with semantic theories of information. If we analyse the Shannon and Weaver equation, as presented by Burgin, and with some changes to its initial posing when considering how to measure the amount of information, we need (in our case) to make reference to the semantic value of the information, to the meaning in terms of what is said, unlike what Burgin poses as information (the capacity to choose between two independent meanings of the significance, with reference to what could be said).

This difference in approach requires a new postulate because it is now a question of knowing to what extent an information retrieval system is behaving better, i.e., of knowing if it is providing relevant information. For this reason we have to redefine the equation, more along the lines expressed by Wagensberg '*entropy contains the disinformation of the macroscopic observer and represents the number of microscopic configurations compatible with the state of equilibrium in question*' (Wagensberg 1994: 30).

For our purposes, faced with the semantic ambiguity contained in a document, there is a number of possibilities of binary options of relevance or non-relevance, determined by the intrinsic value of each separate term in the text of the document. So we would have:

$$\mathbf{S = -I} \text{ as probability, disorder and disinformation}$$

$$\mathbf{- S = I} \text{ as improbability, order and information}$$

So the Shannon and Weaver equation becomes:

$$\mathbf{I} = \mathbf{-\Sigma} p_i \log p_i$$

where p_i is the probability of a term's belonging to a class, which in thermodynamics is known a *degrees of freedom*; in our case we would have i = 1, ..., n classes or terms to establish the canonic representations of the questions, which would be defined by:

$$\mathbf{P}_i = n_{i/N}$$

where N is the number of documents in the collection and n_i is the number of documents in the collection belonging to class i, or relevant to the term i.

In these circumstances, when $N = n_i$ and, consequently $I = 0$ there is no freedom of choice, this means that we have reached certainty, that the response has been 100% correct. For Shannon and Weaver it means that there is no information since there is no freedom of choice between the **m** messages to choose, but in this case we interpret that the retrieval has been completely successful and that it is not necessary to take any decisions of relevance; in short, we are in an environment of data retrieval. Effectively, when we make a a search in a relational database the expectations of success are 100%, Shannon is right, there is no choice, since there is one and only one set of the response (nobody would expect that on obtaining the list of students in a subject from a database that the number of students on the list would be higher or lower than the number of students enrolled).

For the remaining cases where $N > n_i$, then $I > 0$. Thus, the degree of freedom or choice increases and, in consequence, we are talking now of a response success percentage of below 100%, and this success will be lower the greater the value of I. We are in an information retrieval environment ruled by the laws of semantic ambiguity, where a distance is created between the information needs of the user and how to satisfy these through the set of the retrieval response; in short, whether choices of relevance need to be made or not.

Using the above analysis, we can reformulate the problem of interaction between a system and its environment – the retrieval of information to meet user requirements – to establish a postulate for the evaluation. Before explaining the model, the leading players have to be considered. This scenario will be described following Baeza-Yates:

> *an information retrieval model is the quadruple [D, Q, F, R(qi,dj)] where D is a set made up of logical views (or representations) for the documents in the collection. Q is a set made up of logical views (or representations) for the information needs of the user. These representations are called questions. F is a framework for the representation model of documents, questions and their relations. R(q_i,d_j) is an alignment function that associates a real number with a question qi E Q and a representation of the document dj ED. This alignment defines an order among the documents with respect to the questionS q_i. (Baeza-Yates 2011: 58).*

In the first place, among the information flows, collections of documents or primary documents and their representation for storage or logical viewing, there is a differential (understood as the approximation or distance between the real and its representational form), which is the logical view of the complete text, or a set of terms or key words which we call $\mathbf{\Delta}_f$.

information sources $\mathbf{\Delta}_f$**logic view (representation)**

The representation of the information need by a question, be it in a Boolean canonic form or through a number of terms, leads us to define another differential, that we will call $\mathbf{\Delta}_q$.

information need $\mathbf{\Delta}_q$ **logic view (query)**

As indicated earlier, one characteristic of information retrieval systems is the order of the response. This is the consequence of entropy I being greater than zero (i > 0); this differential is tightly linked to the previous one and, hence, some of the existing systems seek to improve their response through specific algorithms (e.g. *PageRank* or *TrustRank*) or by processes refining the search or feedback processes too adjust the order. We will call this differential$\boldsymbol{\delta}_r$

information search $\mathbf{\Delta}_b$ **documents of the response**

Whatever the information retrieval model used for a search (boolean, algebraic or probabilistic), it is clear that another differential will exist, perhaps that which represents most weight, given that the similarity functions are those which decide the crucial issue of retrieval, i.e., when a, document is relevant or not and to what extent it is relevant. We will call this differential $\mathbf{\Delta}_r$

documents of the response$\mathbf{\Delta}_r$**ranking**

In short, we can state that the effectiveness of the information retrieval, the factors that intervene therein and that are responsible for the positive value of the entropy can be represented by the

$$\mathbf{\Delta}_{f} + \mathbf{\Delta}_{q} + \mathbf{\Delta}_{b} + \mathbf{\Delta r} = \textbf{Effectivenes}$$

These four differentials can be grouped as follows:

$$\mathbf{\Delta}_{f} + \mathbf{\Delta}_{q} = \mathbf{\Delta}_{\textbf{human}}$$

$$\mathbf{\Delta}_{b} + \mathbf{\Delta}_{r} = \mathbf{\Delta}_{\textbf{computational}}$$

In effect, although the processes for obtaining logical views (in both texts and questions) are highly automated, it is no less true that the human factor in these extremes is primordial when establishing differences (we should not forget that feedback is a mechanism to enhance the query on the basis if the response). Undoubtedly, the human is a vital issue, as we see in the work of Wilson (2000: 49) who identifies four participating elements:

- Information behaviour, or the whole of human behaviour in relation to information sources and channels, including the active search for information and the use of this (including word of mouth and passive reception).
- Information seeking behaviour, or the intentional search for information (manually or automatically) to meet an information demand or to reach a goal.
- Information searching behaviour, or the micro-level of behaviour used by the searcher in the interaction with all types of information systems. This comprises all the interactions with the system, be they in the sphere of person-machine (using the mouse and links, etc.) or in the intellectual sphere (adopting a Boolean search strategy, establishing criteria for deciding which of the selected books on a shelf is of most use), which also involves intellectual activities like judging th e relevance of the data or the information retrieved.
- Information use behaviour, the behaviour of the physical and mental acts involved in incorporating the information with that already present in the person's knowledge. These could be physical acts such as marking the sections of a text to indicate their importance or relevance, or mental acts that imply, for example, comparing new information with that existing.

It is worth making a brief aside to establish the coherence of the systemic approach to information retrieval systems. We have indicated that information retrieval is an open system. In these systems when we find non linear states, discontinuities and instabilities arise and spontaneous fluctuations may become irreversible, so leading to new states known as *dissipative structures* (in nature these are responsible for biodiversity). This is self-regulation and it is not alien to an information search operation in which the response in its whole gives results that may have nothing to do with the initial search strategy but which may be of great use in satisfying other information needs that arise at the same time. These and other similar situations in which the changing human factor participates serve as a base for Wilson (2000) to propose different activities which describe how the user interacts with the information sources, makes searches within them and uses the information. In all types of behaviour, there is a dependence of the differentials related to the human factors mentioned, insofar as users may through their actions affect the value of these differentials. Thus, informational behaviour and the techniques for representing texts and questions are crucial in improving the response.

The computational aspects have been well described in the previous sections, and the evolution and the most important alternatives in which information retrieval systems develop today are clear. Let us end by analysing information flows and the interactions if, through a similar systemic approach to that of Wagensberg (1985), we were to establish information retrieval as a system that interacts with its environment, i.e., the users. There are four quantities to be measured to establish the balance and, thus, to ascertain why we evaluate information retrieval.

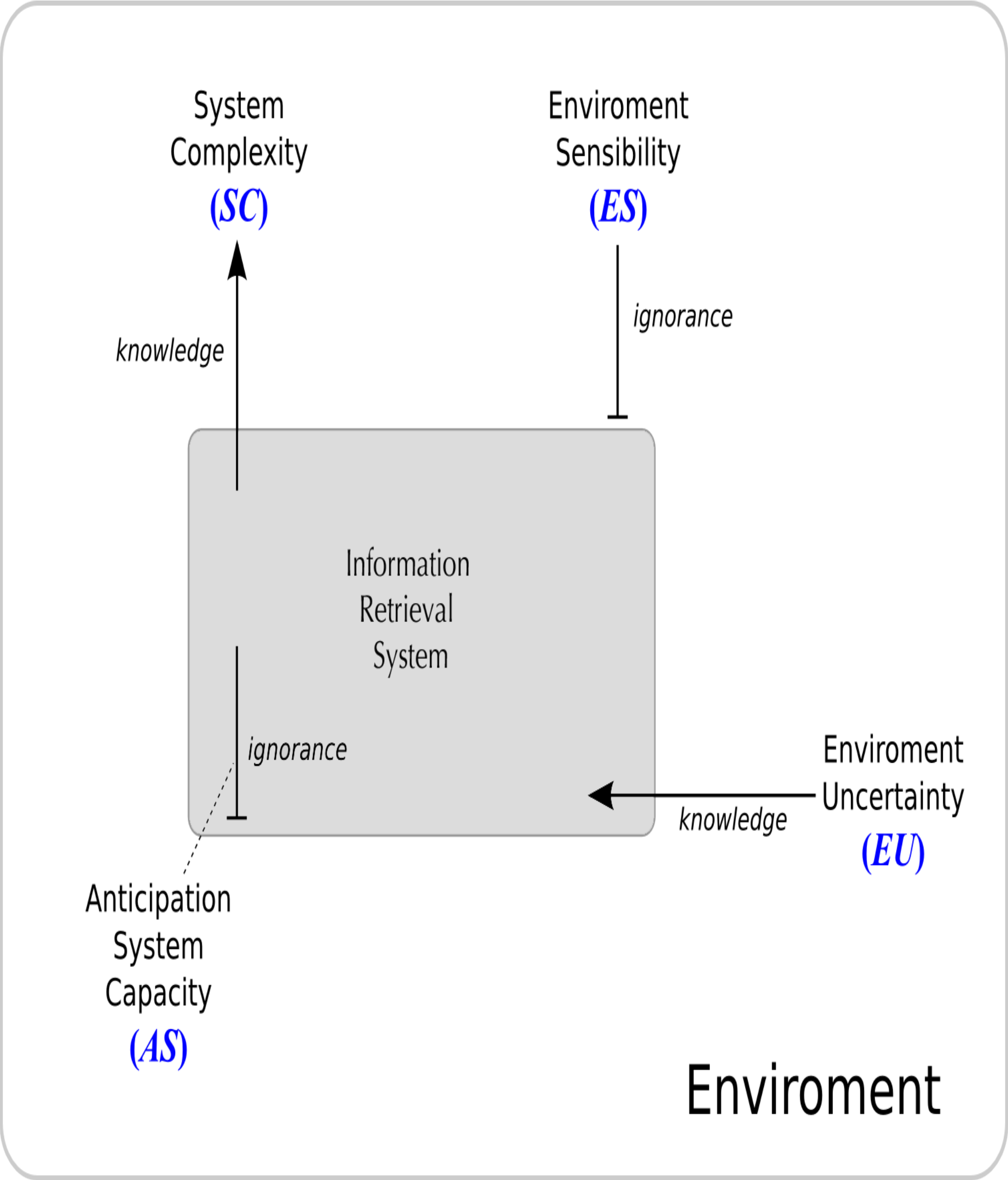

Figure 3: The information retrieval ecosystem.

We consider first the system as a source where the response to the user is originated. The amount of information contained in the source depends on its potential behavioural diversity (key words have fewer accessible states than abstracts, and these have fewer than complete texts); this is the complexity of the system (**SC**). These are the data structures (terms, metadata, thesauruses, indexes, etc.), the thematic content, the readability.

We will now broach the environment as the source of a query to the system. The original information depends here also on the richness of the possible behaviour of the source, in this case the environment (a non specialist user has fewer accessible states than a specialist one, a scientist), it is a question of the environment's complexity or, rather, the environment's uncertainty (**EU**). These are the needs and the behaviour of the users in the form of language.

But not all the information stored in the system reaches the environment, or vice versa. The first error lies in the diversity of behaviour that we can find in a system, once a certain environmental behaviour has been established. The smaller this error, the fewer the doubts of the system regarding its environment and the more the environment limits the system's possibilities (the response of an information retrieval operation changes little for the user who employs single terms in the search and one uses feedback). It is the system's capacity of anticipation (**AS**) that matters, the search engines, the search and alignment algorithms on which the retrieval models are based. The opposite error, i.e., information leaving the environment does not reach the system, depends analogously on the variety of states compatible with the system's behaviour. The lower this value, the more affected the environment will be by the system (a file with bibliographical references has less influence than an information retrieval system on the user when satisfying information requirements). In this case we are dealing with the environment's sensitivity (**ES**), semantic ambiguity, intelligibility, etc. According to Wagensberg:

> *the net information that reaches a destination is obtained by subtracting the error from the source information. Hence, the complexity of the system minus its capacity of anticipation is just the information that the environment's behaviour provides on the system's behaviour. Contrarily, the uncertainty of an environment minus its sensitivity is merely the information that a system's behaviour provides on the environment's behaviour. (Wagensberg 1985:)*

Conrad (1993) establishes (equation 3) that both sets of information contain identical amounts of information, and this equality regulates the changes of any ecosystem. Under theses premises, any movement of a term in the equation implies the immediate re-accommodation of one (or all)of the remaining ones. For example, if the environment uncertainty increases (a complex or ambiguous question), the system would have to increase its complexity by improving a thesaurus's relationships or improving its capacity of anticipation by refining its search and alignment algorithms, or by disqualifying its effect on the environment by incorporating logical inference processes into its information retrieval. In short, this equation provides for four parameters and a principle of equilibrium to be respected. When this happens, the communication between the system and environment is able to combat all the difficulties by not violating this equilibrium so we can say that an adaptation has been produced.

$$\mathbf{SC - AS = EU - ES \ (3)}$$

If an episode occurs in one of the terms and none of the other three terms were able to redress the balance, the adaptation would break down and the system would go into a crisis:

> *the system them would shut down or would brusquely change to another structure, it would organize itself into a clear rebellion*

> *against its environment. A catastrophe (bifurcation) would come about. And the new system could, if it found a way of being compatible with the fundamental newly established identity, be a good system. And the story would go on. It is the same essence of change (Wagensberg, 1985: 49).*

Consequently, we can establish the connections between the four measures and the differentials that govern the effectiveness of information retrieval as:

$$(\mathbf{SC}, \mathbf{AS}) = \boldsymbol{\delta}_{f\,+}\,\boldsymbol{\delta}_{q}$$

$$(\mathbf{EU}, \mathbf{ES}) = \boldsymbol{\delta}_{bb+}\boldsymbol{\delta}r$$

Put another way, **SC** and **SA** will be parameters that are sensitive to the improvements in the representation of documents and to the representation of questions as a consequence of the users' behaviour. On the other hand, **EI** and **ES** will be parameters that are sensitive to the improvements implemented by the information retrieval system and, hence, those of the technological innovations.

Final consideration

We have seen the circumstances in which information retrieval systems have evolved and how they have had to be adapted to the growing demands of the users, particularly the search engines used for the Internet. Some processes have fallen into disuse with the passing of time, and today there are only three great search engines on the market (possibly two, following the alliance in 2012 between Yahoo! and Bing Network. The survivors have had to to make changes to balance this equation and to get rid of the differentials that affect these processes and significantly affect the equilibrium and, hence, adaptation.

With all the instruments and processes they involve, these systems, in short, seek to overcome epistemological chance, that due to our own ignorance (weak algorithms, erroneous choices of terms when searching, badly constructed thesauruses, etc.). We might say that if we distance ourselves from the equilibrium, essentially the data, we are paving the way for the intervention of chance, essentially semantics, and are left at the mercy of the fluctuations of a dissipative environment, and hence only the contributions of genuine novelties could bring any change, essentially search engines, although there would always be ontological chance, that which describes the pure eventuality that acts permanently throughout the universe.

We believe that a formal framework has been established that recognizes not only the factors that participate in information retrieval (in principle, well-known), but one that also establishes how they participate. Thus we now know why we evaluate information retrieval and how to intervene in moments of imbalance.

Acknowledgements

We wish to thank once again and surely not the last, the opportunity provided by Profesor Tom Wilson of participating in this journal, the result of his many years of work and expertise. Without his effort, dedication and understanding this project would not be possible. Thank you very much Tom.

About the authors

Jose-Vicente Rodriguez-Muñoz is Professor at the Department of Information and Documentation of the School of Communication and Information Studies, University of Murcia. He holds a Ph.D. in computer science fron the same institution. He is also Visiting Professor at the University of Havana (Cuba) and Coordinator o the UNITWIN Chair *Information Management in Organizations* sponsored by UNESCO. He is head of the Research Group of Information Technology at the University of Murcia and his research area covers the information management and information storage and retrieval. His an Associate Editor of *Information Research* with responsibility for papers from the Luso-Hispanic regions. He can be reached at jovi@um.es

Francisco-Javier Martínez-Méndez is Lecturer of Information Technology and Dean of the Department of Information and Documentation of the School of Communication and Information Studies, University of Murcia. His PhD is in information science and he is member of the Research Group on Information Technology at the University of Murcia. He teaches and conducts research in information retrieval and strategic information management. He is the author of the blog: IRSWEB, information retrieval on the Web. He can be reached at javima@um.es

Juan-Antonio Pastor-Sanchez is Lecturer of Information Technology and Assistant Dean at the Department of Information and Documentation of the School of Communication and Information Studies, University of Murcia. His research is related to thesauri, ontologies and recently, he has specialized in the semantic Web technology. He can be reached at pastor@um.es

Aladro Vico, E. (2009). *La información determinante*. [Relevant information.] Madrid: Tecnos.

Baeza-Yates, R. & Frakes, W.B. (Eds.) (1992). *Information retrieval: data, structures and algorithms*.Englewood Cliffs, NJ: Prentice Hall.

Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Harlow, UK: Addison-Wesley

Berners-Lee, T. (1989) *Information management: a proposal*. Geneva, Switzerland: CERN. Retrieved 25 November, 2012 from http://www.w3.org/History/1989/proposal.html (Archived by WebCite® at http://www.webcitation.org/6CRBkojo8)

Berners-Lee, T., Hendler, J. & Lassila, L. (2001). The semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, **284**(5), 34-56. Retrieved 25 November, 2012 from http://www.scientificamerican.com/article.cfm?id=the-semantic-web (Archived by WebCite® at http://www.webcitation.org/5hm6NbuPU)

Blair, D.C. (1990). *Language and representation in information retrieval*. Amsterdam: Elsevier Science Publishers.

Burgin, M. (2010). *Theory of information. Fundamentality, diversity and unification*. Singapore: World Scientific Publishing Co.

Conrad, M. (1983). *Adaptability*. New York, NY: Plenum Press.

Cooper, W.S. (1973). On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science*, **24**(1), 87-92

Darwin, C. (1859). *On the origin of the species by means of natural selection, or the preservation of favoured races in the struggle for life*. London: John Murray

De Virgilio, R., Giunchiglia, F. & Tanca, L. (Eds.) (2010). *Semantic Web information management*. Berlin: Springer Verlag.

Foskett, D.J. (1972). A note on the concept of relevance. *Information Storage and Retrieval*, **8**(2), 77-78

Frants, V.I., Shapiro, J. &: Voiskunskii, V.G. (1997). *Automated information retrieval: theory and methods*. San Diego, CA: Academic Press.

Gordon, M. & Pathak, P. (1999). Finding information on the World Wide Web: the retrieval effectiveness of search engines. *Information Processing and Management*, **35**(2), 141-180

Greisdorf, H. (2000). Relevance: an interdisciplinary and information science perspective. *Informing Science*. **3**(2), 67-71

Lancaster, F.W. (1991). *Indexing and abstracting in theory and practice*. London: The Library Association.

Manzelli, P. (1992). *Creativity and science*. Australian Chemistry Resource Book, V.12 128-130

Manzelli, P. (1998). *Creatividad y ciencia: hacia la sociedad del conocimiento*. [Creativity and science: towards a knowledge society] Retrieved 11 March, 2012 from http://www.edscuola.it/archivio/lre/creatividad.html (Archived by WebCite® at http://www.webcitation.org/6CUbLBG3f)

Martínez Méndez, F.J. (2002). *Propuesta y desarrollo de un modelo para la evaluación de la recuperación de información en Internet*. [Proposal and development of a model for the evaluation of information retrieval on the Internet.] Alicante: Biblioteca Cervantes Virtual. (Doctoral thesis, University of Murcia) Retrieved 11 March, 2012 from http://www.cervantesvirtual.com/obra/propuesta-y-desarrollo-de-un-modelo-para-la-evaluacion-de-la-recuperacion-de-informacion-en-internet--0 (Archived by WebCite® at http://www.webcitation.org/6CSdGRyf1)

Martínez Méndez, F.J. & Rodríguez Muñoz, J.V. (2003). Síntesis y crítica de las evaluaciones de la efectividad de los motores de búsqueda en la Web. *[Synthesis and critical evaluation of the effectiveness of Web search engines.] Information Research*, **8**(2), paper 148. Retrieved 18 January, 2012 from http://informationr.net/ir/8-2/paper148.html (Archived by WebCite® at http://www.webcitation.org/5cy4BQPKY)

Martínez Méndez, F.J. y Rodríguez Muñoz, J.V. (2004). Reflexiones sobre la evaluación de los sistemas de recuperación de información: necesidad, utilidad y viabilidad. [Reflections on the evaluation of information retrieval systems: necessity, utility and feasibility] *Anales de Documentación*, No. 7, 153-170. Retrieved 18 January, 2012 from: http://revistas.um.es/analesdoc/article/view/1651/1701 (Archived by WebCite® at http://www.webcitation.org/6CRBwGXWA)

Mizzaro, S. (1997). Relevance: the whole history. *Journal of the American Society for Information Science*, **48**(9): 810–832. Retrieved 5 December, 2012 from http://ilps.science.uva.nl/Teaching/0405/AR/part2/rel-hist-jasis.pdf (Archived by WebCite® at http://www.webcitation.org/6CgqyzPAR)

Mizzaro, S. (1998). *How many relevances in information retrieval?* Udine, Italy: Università degli Studi, Department of Mathematics & Computer Science, Retrieved 18 March, 2012 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.97.9124&rep=rep1&type=pdf (Archived by WebCite® at http://www.webcitation.org/6CSdeQEcn)

Moreiro González, J.A. (2004). *El contenido de los documentos textuales: su análisis y representación mediante lenguaje natural*. [The textual content of documents: analysis and natural language representation.] Gijón, Spain: Trea.

Pastor Sánchez, J.A. (2011). *Tecnologías de la Web Semántica*. [Technologies of the Web.] Barcelona: Universitat Oberta de Catalunya.

Pors, N.O. (2000). Information retrieval, experimental models and statistical analysis. *Journal of Documentation* **56**(1), 71-90.

Salton, G. & Mc Gill, M.J. (1983). *Introduction to modern information retrieval*. New York, NY: McGraw-Hill.

Shannon, C.E. and Weaver, W. (1981). *Teoría matemática de la Comunicación*. [Mathematicsl theory of communication.] Madrid: Ediciones Forja.

Wagensberg, J. (1985). *Ideas sobre la complejidad del mundo*. [Ideas on the complexity of the world.] Barcelona: Tusquets Editores. (Metatemas 9.)

Wilson, T.D. (2000). Human information behavior. *Informing Science*, **3**(2), 49-56. Retrieved 25 November, 2012 from http://inform.nu/Articles/Vol3/v3n2p49-56.pdf (Archived by WebCite® at http://www.webcitation.org/5LCZJnFl2)

Rodriguez-Muñoz, J.V.; Martínez-Méndez, F.J. & Pastor-Sanchez, J.A. (2012). "The ecosystem of information retrieval" *Information Research*, **17**(4) paper 553. [Available at http://InformationR.net/ir/17-4/paper553.html]

Find other papers on this subject

Scholar Search Google Search Bing

Check for citations, using Google Scholar

Like 0 Tweet 3

Contents | **Author index** | **Subject index** | **Search** | **Home**