UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de Marília

Information Retrieval: representation of the subjective

Edberto Ferneda

Como citar: FERNEDA, Edberto. Information Retrieval: representation of the subjective. *In*: JORENTE, Maria José Vicentini; SAN SEGUNDO, Rosa; MONTOYA, José Antonio Frías; MARTÍNEZ-ÁVILA, Daniel; SILVA, Stephanie Cerqueira. (org.). **Digital Curation and Gender in Information Science:** access and preservation. Marília: Oficina Universitária; São Paulo: Cultura Acadêmica; San José: Ediciones Edicic, 2022. p. 155-169. DOI: https://doi.org/10.36311/2022.978-65-5954-328-1.p155-169.

CHAPTER 8

INFORMATION RETRIEVAL: REPRESENTATION OF THE SUBJECTIVE

EDBERTO FERNEDA
Universidade Estadual Paulista

ABSTRACT

Information retrieval designates the operation by which documents are selected from a collection based on a specific informational demand. A document is retrieved if its representation totally or partially coincides with the representation of the user's need. The correct interpretation of such representations is fundamental for an information system efficiency, which involves processes whose formalization and automation are only possible through simplifications of typically subjective concepts. These simplifications directly or indirectly affect the information system efficiency. This work presents and assesses the forms of computational representation of concepts and operations part of the information retrieval process. The automation of the information retrieval process allows to operate large amounts of data in a fast and agile way, but it does not necessarily provide consistent or satisfactory results. Judging the relevance of information implies procedures based on human capacities and abilities of abstraction, apprehension and representation of its meaning.

Keywords: information retrieval, information representation, meaning of information, relevance, subjectivity.

https://doi.org/10.36311/2022.978-65-5954-328-1.p155-169

RESUMEN

La recuperación de información designa la operación mediante la cual se seleccionan los documentos de una colección en función de una demanda determinada de información. Se recupera un documento si su representación coincide total o parcialmente con la representación de la necesidad del usuario. La correcta interpretación de tales representaciones es fundamental para la eficiencia de un sistema de información, que involucra procesos cuya formalización y automatización solo son posibles por medio de simplificaciones de conceptos típicamente subjetivos. Estas simplificaciones afectan directamente en la eficiencia de los sistemas de información. Este trabajo presenta y evalúa las formas de representación computacional de los conceptos y operaciones que hacen parte del proceso de recuperación de información. La automatización del proceso de recuperación de información viabiliza la operación de grandes cantidades de datos de forma rápida y ágil, pero no necesariamente proporciona resultados consistentes o satisfactorios. El juicio de relevancia de la información implica procedimientos basados en las capacidades y habilidades humanas de abstracción, aprehensión y representación de su significación.

Palabras clave: recuperación de información, representación de la información, significación de la información, relevancia, subjetividad.

1 INTRODUCTION

The search for information using some type of technological resource has become a common activity in contemporary society. When we search on the Web, we are looking for information to satisfy a certain need. Information is considered relevant if it brings the knowledge we need at a given time, in a given situation.

The idea of using electronic devices in the search for information had its genesis with the works by Paul Otlet (1934) and later with the article entitled "As We May Think", by Vannevar Bush (1945). The ideas conveyed in these works paved the way for several studies in the following decades. In the early 1950s, mathematician, physicist and computer scientist Calvin Northrup Mooers (1951) created the term "Information Retrieval", inaugurating an area of research that has consolidated and strengthened over the years. The popularization of the Internet and the emergence of the Web have brought new challenges and great interest in research and development of techniques to assist in information search and retrieval in this worldwide collection.

From the first investigations to the present day, the role of information retrieval systems has gone from simple experimental tools to systems for everyday use, useful to everyone who needs information for their activities. During this period, the accelerated technological advance and countless ideas, concepts and techniques were proposed and developed. However, the search for relevant and useful information is still an arduous task. This difficulty leads to reflection on the main elements involved in the information retrieval process, which are apparently alien to technological advances, or at least to the technologies currently available (Ferneda, 2013).

Information retrieval is the operation through which documents from a collection are selected according to a certain information demand. In essence, retrieval occurs through the comparison between document representation and the user's information need representation. A document is retrieved if its representation fully or partially matches the representation of the user's need. Retrieving information therefore implies operating selectively on a set of information items, which involves processes whose formalization and automation are only possible through simplifications of typically subjective concepts (Ferneda, 2003).

The objective of this text is to assess the forms of computational representation of the inherently subjective concepts and operations that are part of the information retrieval process.

2 THE RELEVANCE OF RELEVANCE

The concept of relevance is crucial in Information Retrieval, often used in the statement of the objectives of this area. It is a fundamental issue and a central concern for the functioning and assessment of information retrieval systems (Saracevic, 2017; Mizzaro, 1997; Cooper, 1971).

The term "relevance" is generally used to identify an element that stands out in a given set. It is also used to discriminate an object "of great value or interest", or even to refer to "what matters or is necessary"[1].

The concept of relevance can be expressed by different terms. Vannevar Bush (1945) used the expression "item of momentary importance"; Mooers (1951) referred to "useful information". Terms such as "relevant", "valuable", "useful", "significant" are used in different connotations, but usually with meaning underlying relevance.

> As most fundamental notions, relevance is intuitively well understood – nobody has to explain it to anybody in the world. That is its strength. That is why the systems aiming at retrieval of relevant information to users, including search engines and a variety of search apps in social media, are so well accepted globally – differences in cultures, societies, and mores do not matter. However, relevance is a human, not a technical, notion. That is its weakness. As all human notions, relevance is messy. Relevance encompasses many variables that are hard to control and even fathom formally. Relevance always, repeat always, involves a context as well. All the search algorithms in all the systems in the world are trying to approximate, with various degrees of success, the human notion of relevance. That is what they are all about, that is why they exist. (Saracevic, 2015, p. 27).

[1] FERREIRA, Aurélio Buarque de Holanda. Novo Dicionário da Língua Portuguesa.

Relevance always involves a relationship. There is always a “to” associated with relevance that refers to a context, an issue in question. Something is relevant to someone or to a certain context. The concept of relevance is not necessarily binary, there are degrees that change as intentions and cognitive horizons change, or when the subject in question changes (Saracevic, 2017, p. 17). According to Sperber and Wilson (2005, p. 224, our translation), “intuitively, relevance is not a matter of all or nothing, but a matter of degrees”. Assigning these degrees of relevance is an inherently subjective process.

Saracevic (1975), Swanson (1986) and Harter (1992) distinguish two types of relevance: “objective relevance” and “subjective relevance”. Objective relevance is related to the systems, while subjective relevance is related to the operation and use of such systems by their users. According to Swanson (1986), in an information retrieval system it is up to the user to judge the relevance of the information resulting from a search. This arbitration carries an individual character, a “mental experience” based on each user’s characteristics. Saracevic (2017, p. 24) argues that systems are created by different designers who use different approaches and different development methods. So, in a way, systems are also subjective. Therefore, according to the author, there is no “objective” relevance. Every relevance is subjective, even when formalized in an algorithm.

Similar to the objective-subjective dichotomous classification, several authors use the terminology “system relevance” and “user relevance” (Mizzaro, 1997). The system relevance is a potential relevance that is assumed, defined and formalized from hypotheses or conjectures related to the representation structure of information items, the way these items are organized and the degree of similarity of each item in relation to the search expression. On the Web environment, for example, considering its structure formed by a set of pages connected by links, Google’s basic algorithm (PageRank) starts from the idea that the number of links a web page receives from other pages can serve as a measure of its relevance (Brin & Page, 2012). Library systems use relevance criteria adapted to the representation structure of the items in their collection. The Primo[2] system ranks the results of a search

[2] Ex-Libris Primo is a set of tools developed and marketed by the company Ex-Libris that implement search and retrieval resources for digital object collections.

based on the following relevance criteria (Ex Libris, 2015):

1. **How well an item matches the query.** An item is considered more relevant if the query terms occur in specific metadata fields of the item record (author, title, subject) and if the record's terms appear in the same order as the query;

2. **The academic importance of an item**. The academic significance of the item is calculated from factors unrelated to the query. To calculate the academic importance of an item, it considers whether it was published in a peer-reviewed journal, the number of citations, among other characteristics;

3. **The relevance of an item to the search type.** The system infers the type of search the user is conducting. In a search for a broad or generic subject, the system adds reference articles to its results. In searches for more specific items, the system considers authors, titles or other characteristics to place some items at the top of the results list.

4. **How current an item is.** It is assumed that users generally prefer recent materials.

A retrieval system assigns relevance following criteria formalized by its algorithms. These algorithms have the main function of comparing the representation of each document in the collection with the search expression expressed by the user. The result of this comparison is a numerical value that represents the degree of relevance of each document in relation to the search. This degree of relevance is generally used to order (rank) the set of documents resulting from a search.

From the set of documents resulting from his search, the user judges the relevance of the items retrieved data (user relevance) using their knowledge on the researched subject. Borlund (2003) argues that relevance is a multidimensional cognitive concept whose meaning is largely dependent on users' perceptions and their needs. The user's judgment of relevance is initially based on their need for information. However, the importance given to certain dimensions of relevance can change dynamically, as the user advances in the result analysis.

The concept of relevance has played an important role in information retrieval system development. If the efficiency of a system lies in its ability to retrieve relevant documents, this efficiency can be measured by the proximity between the system relevance and the relevance for the user. The system relevance can be formalized using characteristics related to the item organization and representation in a collection. However, the relevance for the user escapes any kind of formalization or representation.

3 THE REPRESENTATION INCOMPLETENESS

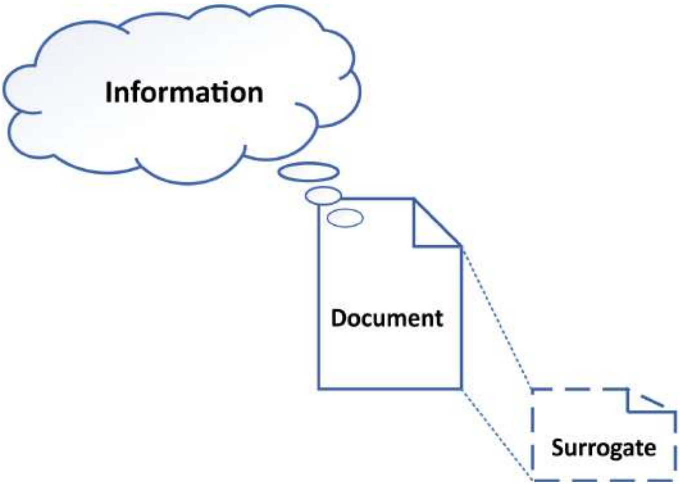
> ...In that Empire, the Art of Cartography reached such Perfection that the Map of a single Province occupied an entire City, and the map of the Empire, an entire Province. Over time, these Huge Maps were not satisfactory and the Colleges of Cartographers put up a Map of the Empire that had the size of the Empire and punctually coincided with it. Less accustomed to the study of cartography, the following generations understood that this dilated Map was useless and not without impiety, they delivered it to the inclemencies of the Sun and Winters. In the Western deserts, the shattered ruins of the map remain, inhabited by animals and beggars; in the whole country there is no other relic of the Geographical Disciplines. (Suárez Miranda, Viajes de varones prudentes, book four, chapter. XLV, Lérida, 1658, our translation).
>
> Jorge Luis Borges – from rigor in science

Every representation is incomplete. If it weren't incomplete, it wouldn't be a representation. A representation is usually shorter or briefer than the represented object, restricted to the considered most relevant characteristics. Therefore, creating a representation involves choices about what will be included in it and what will be discarded. Something of the original is always lost. A representation will always be a distorted version of the real, even if only for its incompleteness (Saracevic, 1991).

The information retrieval process involves two representation instances: the representation of each information item from a given collection and the representation of the user's information need through a search expression (query). According to Belkin, Oddy and Brooks (1982a), these two representations are distinct in nature. An information item (document) is the representation of a "coherent state of knowledge", while a query is the representation of an "anomalous state of knowledge". There are situations where the user is able to specify exactly what information is needed to solve a given problem. However, the most common situation is when the user does not have prior knowledge of the information he/she needs, nor is he/she able to formalize it in a search expression.

The representation of a document and includes the descriptive elements that identify and characterize it in a collection, as well as the elements indicative of its informative content. Figure 1 illustrates the elements of the document representation process defined by Mizzaro (1998).

Figure 1: Representation of the document representation process



Source: by the author

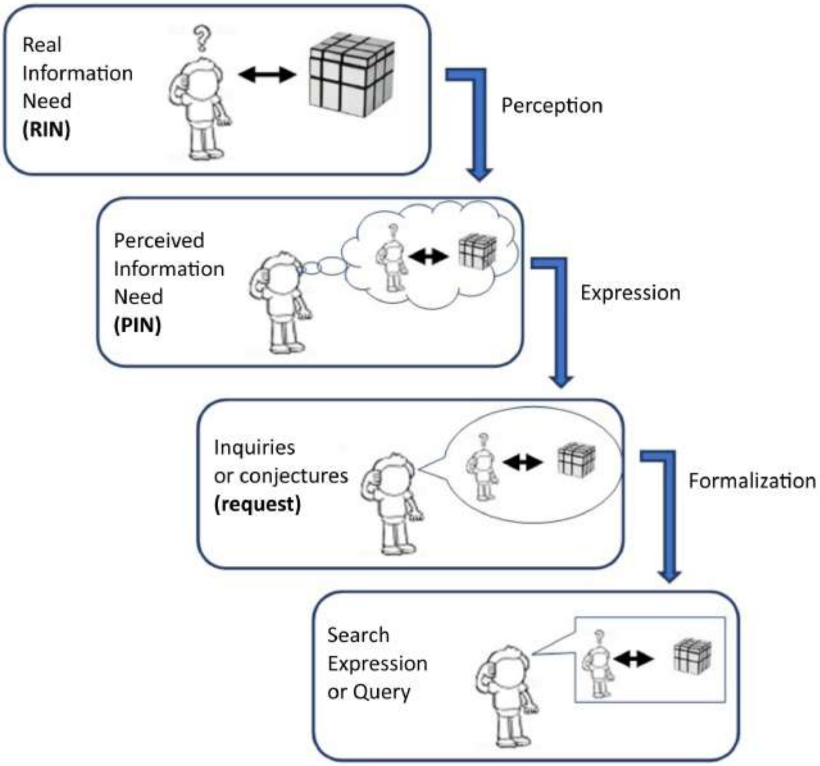
A document is the physical representation of knowledge, the materialization of information. It is the entity the user of an information retrieval system obtains in response to his/her search. The "surrogate" is the representation of the document, consisting of elements that distinguish it from other items in the collection. It is the element that will be compared with the search expression, responsible for retrieving the document. Mizzaro (1998) orders these three elements as follows:

surrogate < document < information

In an information retrieval system, the documentary collection is constituted a priori, and can be processed by automated techniques such as automatic indexing, text mining, among others. On the other hand, the user's need for information is only perceived after its enunciation through a search expression and its interpretation is hampered by the reduced number of terms that are normally used. However, from its definition, the search expression can be used in interactive processes that aim to resolve possible ambiguities or that allow its semantic enrichment (Pansani Junior, 2021).

Figure 2 illustrates the process of representing a search defined by Mizzaro (1998). It has four entities

Figure 2: Representation of the search for information process



Source: adapted from Mizzaro (1998).

According to Belkin, Oddy and Brooks (1982a; 1982b), an information need arises from a recognized anomaly in the user's state of knowledge about some issue or problematic situation he/she cannot precisely specify what is necessary to solve it. Mizzaro (1998) calls this initial need the Real Information Need (RIN). The user realizes his/her need and builds a mental representation, possibly incomplete or incorrect in relation to the RIN: the Perceived Information Need (PIN). Through conjectures or inquiries (request), the user expresses his/her need in a human language, a natural language. Finally, the user formalizes his/her inquiries in a query using the language provided by the information retrieval system. At each representation level there is a loss or distortion in relation to the previous level. The constituent elements of this process can be ordered as follows (Mizzaro, 1998):

query < request < PIN < RIN

Therefore, the query is the possibly incomplete linguistic materialization of an information need, after a sequence of mental representations.

In essence, the information retrieval process is carried out by comparing representations: the user's information need representation and each document representation in a collection. The result of this comparison will usually be a number that represents the document relevance degree in relation to the search and will position it in the list of results.

4 MATHEMATICAL INACCURACY

> [...]
> - "All right," said Deep thought. - The Answer to the Great Question...
> - "Yes...!"
> - "Of Life, the Universe and Everything ..." said Deep Thought.
> - "Yes!"
> - "Yes..." said Deep Thought, and paused.
> - "Yes...!"
> - "Is..."
> - "Yes...!!!...?"
> - "Forty two," said Deep Thought, with infinite majesty and tranquility.
>
> Douglas Adams – The Hitchhiker's Guide to Galaxies

The first computers weighed several tons and their programming was done by directly connecting their circuits. In the 1950s, programming was done by transmitting instructions in binary code through cards or punched tapes. With the emergence of programming languages, binary code was restricted to the core of the computer and communication with the outside world was done by a new program layer. "What was once an interface becomes an internal organ" (Lévy, 1993, p. 101, our translation). Currently, computers are made up of a set of devices and program layers that communicate with each other, allowing a great distance from their binary core.

> Binary, informatics? Undoubtedly, at a certain level of its functioning, but it has been a while since most users have no relationship with this interface. How is a hypertext or a drawing program "binary"? (Lévy, 1993, p. 102, our translation).

In response to the question posed by Pierre Lévy, we can confirm that we currently use a computer without knowledge on how its circuits work, in the same way as we use any other electronic device. However, a computer's binary soul cuts through all its program layers and limits its ability to perform tasks that most humans do with relative ease.

In the information retrieval process, computational resources enable the operation of large document collections, such as the Web. However, the nature of computers requires the mathematization of typically subjective concepts and processes. Relevance, now stripped of its subjectivity, becomes a number. The primary strategy for automating the document representation process (indexing) is simple word count. The words with the highest number of occurrences on the surface of the textual content of a document are elected as representatives of its intellectual content. The need for information is represented by a set of words devoid of their meanings.

The automation of the information retrieval process imposes a logic in which the information must be numerically defined within a closed system, which disregards some human factors involved in this process.

5 CONSIDERATIONS

The term “subjective” is defined as “what belongs to the thinking subject and to his/her innermost being”; “pertaining to or characteristic of an individual; individual, personal, particular”[3]. Subjective is everything that is proper to the subject or relative to it. It is something that is based on an individual interpretation.

The concepts involved in the information retrieval process are typically subjective. The development of computer systems requires simplifications of such concepts so that it is possible to formalize and represent them through algorithms and programs. These simplifications directly or indirectly affect the efficiency of information systems. We have observed that most of the research in Information Retrieval is focused on the search for more efficient ways to represent the subjectivity involved in this process.

The automation of the information retrieval process enables the operation of large amounts of data in a fast and agile way. However, it does not necessarily provide consistent or satisfactory results. Information, considered in its common- sense connotation, is directly related to its meaning, which implies procedures based on human capacities and abilities of abstraction, apprehension and representation of its meaning.

REFERENCES

Belkin, N. J., Oddy, R. N., & Brooks, H. M. (1982a). ASK for Information Retrieval: Part I. Background and Theory. *Journal of Documentation*, *38*(2), 61-71.

Belkin, N. J., Oddy, R. N., & Brooks, H. M. (1982b). ASK for information retrieval: Part II. Results of a design study. *Journal of Documentation*, *38*(3), 145-164.

Borlund, P. (2003). The concept of relevance in IR. *Journal of the American Society for Information Science and Technology*, *54*(10), 913–925.

[3] FERREIRA, Aurélio Buarque de Holanda. Novo Dicionário da Língua Portuguesa.

Brin, S. & Page, L. (2012). Reprint of: The anatomy of large-scale hypertextual Web search engine. *Computer Networks*, *56*(18), 3825–3833.

Bush. V. (1945). As We May Think. *Atlantic Monthly*, 176, 101–108.

Cooper, W.S. (1971). A Definition of Relevance for Information Retrieval. *Information Storage and Retrieval*, 7, 19-37.

Ferneda, E. (2003). *Recuperação da Informação: análise sobre a contribuição da Ciência da Computação para a Ciência da Informação* (Tesis doctoral). Escola de Comunicação e Artes, Universidade de São Paulo, São Paulo, SP, Brasil.

Ferneda, E. (2013). *Ontologia como recurso de padronização terminológica em um Sistema de Recuperação de Informação* (Relatório de Pós-Doutorado). Programa de Pós-Graduação em Ciência da Informação, Universidade Federal da Paraíba. João Pessoa, PB, Brasil.

Ex Libris. (2015). Primo Discovery: Search, Ranking, and Beyond.

Harter, S. P. (1992). Psychological relevance and information science. *Journal of the American Society for Information Science*, *53*(4), 257–270.

Mizzaro, S. (1997). Relevance: The Whole History. *Journal of the American Society for Information Science*, *48*(9), 810-832.

Mizzaro, S. (1998). How many relevances in Information Retrieval? *Interacting with Computers*, *10*(3), 303–320.

Mooers, C. N. (1951). Zatocoding applied to mechanical Organization of Knowledge. *American Documentation*, *2*(1), 20-32.

Otlet, P. (1934). *Traité de documentation: le livre sur le livre - théorie et pratique*. Bruxelles: Mundaneum.

Pansani, E. A., Jr. (2021). *Contextualização e Expansão de Consultas em Sistemas de Recuperação de Informação: um método baseado em ontologias de domínio* (Tesis doctoral). Faculdade de Filosofia e Ciências, Universidade Estatual Paulista-UNESP, Marília, SP, Brasil.

Lévy, P. (1993) *As tecnologias da Inteligência: o futuro do pensamento na era da informática*. São Paulo: Editora 34.

Saracevic, T. (1975). Relevance: A review of and framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, *26*(6), 321-343.

Saracevic, T. (2017). *The notion of relevance in Information Science: everybody knows what relevance is. But, what is it really?* Williston: Morgan & Claypool.

Saracevic, T. (1991). Information science: origin, evolution and relations. En International Conference on Conceptions of Library and Information Science: historical, empirical and theoretical perspectives, 1991, Finland. Proceedings... Helsink.

Sperber, D., & Wilson, D. (2005). Teoria da relevância. *Linguagem em (Dis)curso* 5, 221-268.

Swanson, D. R. (1986). Subjective versus objective relevance in bibliographic retrieval systems. *Library Quarterly*, *56*(4), 389-398.