Journal of Library Services and Technologies, *2*(2), 48-56, June 2020
ISSN: 2616-1354 (Print) 2636-7424 (Online) Available online at credencepressltd.com
DOI: 10.47524/jlst.v2i2.5

Current trends in information retrieval systems: review of fuzzy set theory and fuzzy Boolean retrieval models

Jonathan N. Chimah, *PhD*
Ebonyi State University Library,
Abakaliki, Nigeria
E-mail: jonachim2000@yahoo.com

Friday Ibiam Ude
Ebonyi State University Library
Abakaliki, Nigeria

Abstract
This paper reviews the concept and goal of Information Retrieval Systems (IRSs). It also explains the synonymous concepts in Information Retrieval (IR) which include such terms as: imprecision, vagueness, uncertainty, and inconsistency. Current trends in IRSs are discussed. Fuzzy Set Theory, Fuzzy Retrieval Models are reviewed. The paper also discusses extensions of Fuzzy Boolean Retrieval Models including Fuzzy techniques for documents' indexing and Flexible query languages. Fuzzy associative mechanisms were identified to include:(1) fuzzy pseudothesauri and fuzzy ontologies which can be used to contextualize the search by expanding the set of index terms of documents;(2) an alternative use of fuzzy pseudothesarui and fuzzy ontologies is to expand the query with related terms by taking into account their varying importance of an additional term and (3) fuzzy clustering techniques, where each document can be placed within several clusters with a given strength of belonging to each cluster, can be used to expand the set of the documents retrieved in response to a query. The paper concludes by recommending that in an electronic library environment, the librarians and information scientists should acquaint themselves with these terms in order to be more equipped in helping library users retrieve online documents relevant to their information needs.

Keywords: Information retrieval systems, Document delivery, Fuzzy Set theory, Fuzzy Boolean retrieval models

Introduction

Information Retrieval System (IRS) came into being as a means of ensuring that information generated and recorded do not get over time. Before knowledge became recorded, individuals formed the repository of knowledge. With libraries, repository of knowledge began to change into recorded form. With the quantity of new information being generated is such that no individual can hope to cope with this information explosion and at the same time make them available to users. This led to the use of information retrieval with minimum cost in time, labour and money. Information retrieval, according to Unagha (2010), is the process of searching some collections of documents in order to identify those documents which deal with a particular subject. Reitz (2004) defined information retrieval as the process, methods and procedures used to selectively recall recorded information from a file of data.

In libraries, searches are made typically for a known item or for information on a specific subject, and the file is usually a human readable catalogue or index, or a computer-based information storage and retrieval system, such as an on-line catalogue and bibliographic database. In the design of such systems, a balance must be attained to facilitate this literature searching activity may legitimately be called an information retrieval system. The

catalogue, index and bibliography, abstract as well as the computer are known as information retrieval systems.

Automated information retrieval systems are used to reduce what has been called information overload. An IR system is a software system that provides access to books, journals and other documents; stores and manages those documents. Web search engines are the most visible IR applications. An information retrieval process begins when a user enters a query into the system. Queries are formal statements of information needs, for example search strings in web search engines. In information retrieval a query does not uniquely identify a single object in the collection. Instead, several objects may match the query, perhaps with different degrees of relevancy.

An object is an entity that is represented by information in a content collection or database. User queries are matched against the database information. However, as opposed to classical SQL queries of a database, in information retrieval the results returned may or may not match the query, so results are typically ranked. This ranking of results is a key difference of information retrieval searching compared to database searching (Jansen and Rieh 2010).

Depending on the application the data objects may be, for example, text documents, images, audio, mind maps or videos. Often the documents themselves are not kept or stored directly in the IR system, but are instead represented in the system by document surrogates or metadata. As Frakes and Baeza-Yates (1992) had noted, most IR systems compute a numeric score on how well each object in the database matches the query, and rank the objects according to this value. The top ranking objects are then shown to the user. The process may then be iterated if the user wishes to refine the query.

Based on this backdrop, this paper examines concepts and the goal of information retrieval systems, current trends in information retrieval systems, explains synonymous concepts (such as imprecision, vagueness, uncertainty, and inconsistency), reviews the fuzzy set theory and its concomitant Boolean retrieval models.

Concept and goal of information retrieval systems

Information retrieval (IR) is concerned with the storage, organization, and searching of collections of information. It has been part of significant part of human technological development since the development of writing. The earliest IR systems were the organization schemes of ancient archives and libraries, such as early Sumerian archives, or the “Pinakes” developed by Callimachus for the library of Alexandria. In the twentieth century the largest impetus to development of automated IR systems was the need to manage increasing larger quantities of information in business and scientific development. Early attempts at automating search capabilities for document collections involved techniques based on punched cards, as well as machines using optical sensing of codes on microfilmed documents (Buckland 2006).

According to Larson (2018), the goal of any IR system is to select the information items (texts, images, videos, etc. which we will refer to as “documents”) that are expected to be relevant for a given searcher (or user) from a large collections of such items. Today these collections range from small sets of items on an individual’s personal computer to the vast resources of the World Wide Web. In all cases the task is the same: to extract some set of items that searchers *wants to have* from all of those they *do not want*. This is not a simple task,

and involves not only the technical aspects of constructing a system to perform such selection, but also aspects of psychology and user behaviour to understand what differentiate the desired items from the non-desired from the particular user's point of view.

The synonymous concepts in information retrieval

The terms imprecision, vagueness, uncertainty, and inconsistency are very often used as synonymous concepts. Nevertheless when they are referred to qualify a characteristic of the information they have a distinct meaning (Motro 1995). Since IR has to do with information, understanding the different meanings of imprecision, vagueness, uncertainty, and inconsistency allows to better understanding the perspectives of the distinct IR models defined in the literature.

Kraft, Bordogna and Pasi (2018) noted that vagueness and imprecision are related to the representation of the information content of a proposition. For example, in the information request, "find *recent* scientific chapters dealing with the *early* stage of infectious diseases by HIV," the terms *recent* and *early* specify vague values of the publication date and of the temporal evolution of the disease, respectively. The publication date and the phase of an infectious disease are usually expressed as numeric values; their linguistic characterization has a coarser granularity with respect to their numeric characterization. Linguistic values are defined by terms with semantics compatible with several numeric values on the scale upon which the numeric information is defined. Imprecision is just a case-limit of vagueness, since imprecise values have a full compatibility with a subject of value of the numeric reference scale.

There are several ways to represent imprecise and vague concepts. One the ways is indirectly, by defining similarity or proximity relationships between each pair of imprecise and vague concepts. If we regard a document as an imprecise or vague concepts, i.e., as bearing a vague content, a numeric value computed by a similarity measure can be used to express the closeness of any two pairs of documents. This is the way of dealing with the imprecise and vague document and query contents in the vector space model of Information retrieval. In this context the documents and the query are represented as points in a vector space of terms and the distances between the query and the documents points are used to quantify their similarity.

Uncertainty is related to the truth of a proposition, intended as the conformity of the information carried by the proposition with the considered reality. Linguistic expressions such as "probably" and "itis possible that" can be used to declare a partial lack of knowledge about the truth of the stated information.

Furthermore, Kraft, Bordogna and Pasi (2018) noted that there are cases in which information is affected by both uncertainty and imprecision or vagueness. For example, consider the proposition "probably document d is relevant to query q." However, the same information content can be expressed by choosing a trade-off between the vagueness and the uncertainty embedded in a proposition. For example, one can express the content of the previous proposition by a new one "document d is more or less relevant to query q." in this latter proposition, the uncertain term probably has been eliminated, but the specificity of the vague term relevant has been reduced. In point of fact, the term more or less relevant is less specific than the term relevant. A dual representation eliminate imprecision and augment the uncertainty,

like in the expression "it is not completely probable that document d fully satisfies the query q".

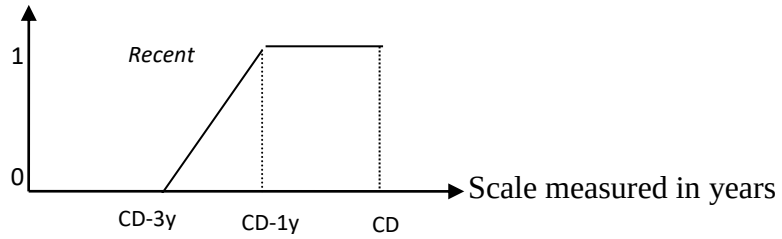

Fig. 1 Semantics of the term "recent" referring to the publication date of a scientific chapter. CD = current date; y = years. **Source**: Kraft, Bordogna & Pasi (2018)

On the basis of what has been said about the trade-off between uncertainty and vagueness to express the same information content, there are two alternative ways to model the IR activity. One possibility is to model the query evaluation mechanism as an uncertain decision process. Here the concept of relevance is considered binary (crisp) and the query evaluation mechanism computes the probability of relevance of a document of *d* to a query*q*. Such an approach, which does model the uncertainty of the retrieval process, has been introduced and developed by probabilistic IR models (Crestani, et al 1998). Another possibility is to interpret the query as the specification of soft "elastic" constraints that the representation of a document can satisfy to an extent, and to consider the term *relevant* as a gradual (vague) concept. This is the approach adopted in fuzzy IR models (Bordogna & Pasi 2000). In this latter case, the decision process performed by the query evaluation mechanism computes the degree of satisfaction of the query by the representation of each document.

This satisfaction degree, called the retrieval status value (RSV), is considered as an estimate of the degree of relevance (or is at least proportional to the relevance) of a given document with respect to a given user query. An RSV of 1 implies maximum relevance; an RSV value of 0 implies absolutely no relevance. And, an RSV value in the interval [0, 1] implies an intermediate level or degree of relevance. For example, an RSV value of 0.5 could imply an average degree or relevance (Kraft, Bordogna and Pasi, 2018).

Inconsistency comes from the simultaneous presence of contradictory information about the same reality. An example of inconsistency can be observed when submitting the same query to several IRSs that adopt different representations of documents and produce different results. This is actually very common and often occurs when searching for information over the Internet using different search engines. To solve this kind of inconsistency, some fusion strategies can be applied to the ranked lists each search engine produces. In fact, this is what metasearch engines do (Bordogna, Pasi & Yager, 2003).

Current trends in information retrieval systems

Some of the current trends in Information Retrieval (IR) research run the gamut in terms of expanding the discipline both to incorporate the latest technologies and to cope with novel necessities. In terms of novel necessities, with the diffusion of the Internet and the heterogeneous characteristics of users of search engines, which can be regarded as the new frontier of IR, a new central issue had arisen, generally

known as the semantic web (Kraft, Bordogna & Pasi, 2018). It mainly consists in expanding Information Retrieval Systems (IRSs) with the capability to represent and manage the semantics of both user requests and documents so as to be able to account for user and document contexts. This need becomes urgent with cross-language retrieval, which consists in expressing queries to search engines. Cross language retrieval not only implies new works on text processing, e.g., stemming conducted on a variety of languages, new models of IR such as the development of language models, but also the ability to match terms in distinct languages at a conceptual level, by modeling their meaning.

Another research trend, according to Kraft, Bordogna and Pasi (2018), is motivated by the need to manage multimedia collections with non-print audio elements such as sound, music, and voice, and video elements such as images, pictures, movies, and animation. Retrieval of such elements can include consideration of both metadata and content-based retrieval techniques. The definition of new IRSs capable to efficiently extract content indexes from multimedia documents, and to effectively retrieve documents by similarity or proximity to a query by example so as to fill the semantic gap existing between low-level syntactic index matching and the semantics of multimedia document and query are still to come.

In addition, modern computing technology, including storage media, distributed and parallel processing architectures, and improved algorithms for text processing and for retrieval, has an effect on IRSs. For example, improved string searching algorithms have improved the efficiency of search engines. Improved computer networks have made the Internet and the World Wide Web a possibility. Intelligent agents can improve retrieval in terms of attempting to customize and personalize it for individual users. Moreover, great improvements have been made in retrieval systems interfaces based on human-computer interface research.

These novel research trends in IR are faced by turning to technologies such as natural language processing, image processing, language models, artificial intelligence, and automatic learning. Also fuzzy set theory can play a crucial role to define novel solutions to these research issues since it provides suitable means to cope with the needs of the semantic web (Sanchez, 2006) i.e., to model the semantic of linguistic terms so as to reflect their vagueness and subjectivity and to compute degrees of similarity, generalization, and specialization between their meanings.

Fuzzy set theory

The notion of a fuzzy set is an extension to normal set theory (Zadeh 1965). According to him, a set is simply a collection of objects. A fuzzy set (more properly called a fuzzy subset) is a subset of a given universe of objects, where the membership in the fuzzy set is not definite. For example, consider the idea of a person being middle-aged. If a person's age is 39, one can consider the imprecision of that person being in the set of middle-aged people. The membership function, μ, is a number in the interval [0,1] that represents the degree to which that person belongs to that set. Thus, the terms *recent* and *early* can be defined as fuzzy subsets, with the membership functions interpreted as compatibility functions of the meaning of the terms with respect to the numeric values of the reference (base) variable. In Fig. 1, the compatibility function of the term *recent* is presented with the numeric values of the time-scale measured in years. Note that here a chapter that has a publication date of the current year or 1 year previous is perfectly *recent*; however, the extent to which a chapter remains *recent* declines steadily

over the next 2 years until chapters older than 3 years have no sense of being recent.

Fuzzy retrieval models

Fuzzy retrieval models have been defined in order to reduce the imprecision that characterizes the Boolean indexing process, to represent the user's vagueness in queries, and to deal with discriminated answers estimating the partial relevance of the documents with respect to queries. Extended Boolean models based on fuzzy set theory have been defined to deal with one or more of these aspects (Bordogna & Pasi, 1995).

It has been speculated that Boolean logic is passé, out of vogue. Yet, researchers have employed p-norms in the vector space model or Bayesian inference nets in the probabilistic model to incorporate Boolean logic into those models. In addition, the use of Boolean logic to separate a collection of records into two disjoint classes has been considered, e.g., using the one-clause-at-a-time (OCAT) methodology (Sanchez, Triantaphyllou & Kraft 2003). Moreover, even now retrieval systems such as Dialog and Web search engines such as Google allow for Boolean connectives. It should come as no surprise, therefore, to see extensions of Boolean logic based upon fuzzy set theory for IR.

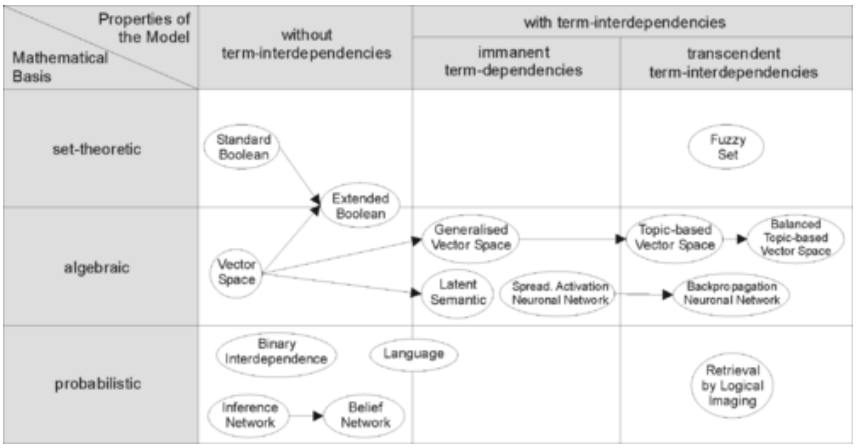

Fig. 2: *Categorization of IR-models. Source: DominikKuropka(2004).*

Extensions of fuzzy Boolean retrieval models

The fuzzy retrieval models have been defined as generalizations of the classical Boolean model. These allow one to extend existing Boolean IRSs without having to redesign them. This was first motivated by the need to be able to produce proper answers in response to the queries. In essence, the classical Boolean IRSs apply an exact match between a Boolean query and the representation of each document. This document representation is defined as a set of index terms. These systems partition the collection of documents into two sets, the retrieved documents and the rejected (non-retrieved) ones. As a consequence of this crisp behaviour, these systems are liable to reject useful items as a result of too restrictive queries, as well as to retrieve useless material in reply to queries (Salton & McGill, 1983).

Fuzzy techniques for documents' indexing

The aim is to provide more specific and exhaustive representations of each document's information content. This means

improving these representations beyond those generated by existing indexing mechanisms.

Flexible query languages

There are query languages that are more expressive and natural than classical Boolean logic. This is defined in order to capture the vagueness of user needs as well as to simplify user-system interaction. This has been pursued with two different approaches. There has been work on the definition of soft selection criteria (soft constrains), which allow the specification of the different importance of the search terms. Query languages based on numeric query term weights with different semantics have been first proposed as an aid to define more expressive selection criteria (Cater & Kraft, 1987).

Fuzzy associative mechanisms

In their work on fuzzy theory, Kraft, Bordogna and Pasi (2018) explained that these associative mechanisms allow to automatically generating fuzzy pseudothesarui, fuzzy ontologies, and fuzzy clustering techniques to serve three distinct but compatible purposes. First, fuzzy pseudothesauri and fuzzy ontologies can be used to contextualize the search by expanding the set of index terms of documents to include additional terms by taking into account their varying significance in representing the topics dealt with in the documents, the degree of significance of these associated terms depends on the strength of the associations with a document's original descriptors. Second, an alternative use of fuzzy pseudothesarui and fuzzy ontology is to expand the query with related terms by taking into account their varying importance of an additional term is dependent upon its strength of association with the search terms in the original query. Third, fuzzy clustering techniques, where each document can be placed within several clusters with a given strength of belonging to each cluster, can be used to expand the set of the documents retrieved in response to a query. Documents associated with retrieved documents, i.e., in the same cluster, can be retrieved. The degree of association of a document with the retrieved documents does influence its RSV. Another application of fuzzy clustering in IR is that of providing an alternative way, with respect to the usual ranked list, of presenting the results of a search.

Conclusion

In the library today, instead of the individual memory, we have the corporate memory – the library catalogues, bibliographies, indexes and computers. These information retrieval systems (tools) contain the bibliographical details of the documents such as the author, edition, call-number, publisher, place of publication, date, etc. The concept and the goal of information retrieval systems have been reviewed. Current trends in information retrieval systems have been highlighted. Attempts have been made to demystify the seeming confusing synonymous concepts in information retrieval which includes imprecision, vagueness, uncertainty, and inconsistency. Related literature has been reviewed on fuzzy set theory, fuzzy retrieval models, extensions of fuzzy Boolean retrieval models, and fuzzy associative mechanisms. It is recommended that in an electronic library environment, the librarians and information scientists should acquaint themselves with these terms in order to be more equipped in helping library users retrieve online documents relevant to their information needs. This would further enhance the utilization of our institutions electronic libraries.

References

Bordogna, G. & Pasi, G. (1995). Linguistic aggregation operators in fuzzy information retrieval. *International Journal of Intelligent Systems,10*(2), 234-248.

Bordogna, G. & Pasi, G. (2000). The application of fuzzy set theory to model information retrieval. In Soft Computing in Information Retrieval: Techniques and Applications. In *Crestani, F.; Pasi, G. Eds. Pysica-Verlag: Heidelberg, Germany.*

Bordogna, G.; Pasi, G. &Yager, R. (2003). Soft approaches to information retrieval on the WEB.*International Journal of Approximate Reasoning.* 2003, 34, 105-120.

Buckland, M. K. (2006). *Emanual Goldberg and His Knowledge Machine*. Libraries Unlimited: Westport, CT.

Cater, S.C. & Kraft, D. H. (1987) TIRS: A topological information retrieval system satisfying the requirements of the Waller-Kraft wish list In *Proceedings of the Tenth Annual ACM/SIGIR International Conference on Research and Development in Information Retrieval.* New Orleans, LA June, 171-180.

Crestani, F.; Lalmas, M.; van Rijsbergen, C.J.; Campbell, I. (1998). Is this document relevant? Probably, *ACM Computer Survey,* 30(4) 528-552.

Frakes, W. B. &Baeza-Yates, R. (1992). *Information retrieval data structures & algorithms*. Prentice-Hall, Inc. . Archived from the original on 2013-09-28.

Jansen, B. J. & Rieh, S. (2010). The Seventeen Theoretical Constructs of Searching and Information Retrieval. *Journal of the American Society for Information Sciences and Technology*, 61(8), 1517-1534.

Kraft, D., Bordogna, G., Pasi, G. (2018). Fuzzy set theory. In: *Encyclopedia of library & information sciences, 4th Edition.* John D. McDonald & Michael Levine-Clark (eds.). Taylor & Francis, pp.1618-1622.

Kuropka, D. (2004). Modellezur Repräsentation natürlichsprachlicher Dokumente. Ontologie-basiertes Information-Filtering und-Retrieval mitrelationalen Datenbanken. *Advances in Information Systems and Management Science, Bd. 10*. Retrieved from: https://www.logos-verlag.de/cgi-bin/engbuchmid?isbn=0514&lng=eng&id=

Larson, R. R. (2018). Information Retrieval System. In: John D. McDonald & Michael Levine-Clark (eds.) *Encyclopedia of Library and Information Sciences*, 4th edition. Taylor & Francis. p.2199.

Motro, A. (1995). Imprecision and uncertainty in database systems. In *Fuzziness in Database Management Systems; Bosc, P., Kacprzyk, J.,Eds.*: Physica-Verlag: Heidelberg, Germany, 3-22.

Reitz, J. M. (2004). *Dictionary of library and information science*, West, Connecticut: Libraries Unlimited.

Salton, G. & McGill, M.J. (1983). *Introduction to Modern Information Retrieval*, New York: McGraw-Hill

Sanchez, E. 2006). *Fuzzy Logic and the Semantic Web*. Elsevier: Amsterdam, the Netherlands.

Sanchez, S.N.; Triantaphyllou, E. & Kraft, D. H. (2003). A feature mining based approach for the classification of text documents into disjoint classes. *Information Processing Management 38*(4), 583-604.

Unagha, A. O. (2010) *Knowledge organization and information retrieval*. Okigwe: Whytem Publishers Nig.

Zadeh, L.A. (1965). Fuzzy sets. *Information Control*, *8.* 338-353.

Dr. Jonathan N. Chimah is the University Librarian of Ebonyi State University Abakaliki and also a Senior lecturer at the Department of Library & Information Science. He is a member of the Nigerian Library Association (NLA), Nigeria Library Information Science Educators (NALISE) and a certified librarian with Librarians' Registration Council of Nigeria (LRCN). E-mail: jonachim2000@yahoo.com; Cell: +2348037976028.

Friday Ibiam Ude, is the immediate past University Librarian of Ebonyi State University. He is currently Faculty of Agriculture & Natural Resources Librarian of EBSU and also a lecturer at the Department of Library and Information Science, Ebonyi State University Abakaliki. He is a member of Nigerian Library Association (NLA) and a certified librarian with Librarians' Registration Council of Nigeria (LRCN).
E-mail: ibiamude7@gmail.com;
Cell: +2347038852971.