Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 201 (2022) 777–782

Procedia Computer Science

www.elsevier.com/locate/procedia

The 3rd International Workshop on Statistical Methods and Artificial Intelligence (IWSMAI'22)
March 22-25, 2022, Porto, Portugal

A Review on recent research in information retrieval

S. Ibrihich[a,*], A.Oussous[b], O. Ibrihich[a], M. Esghir[a]

[a]*Mohammed V University in Rabat, Faculty of Sciences, Laboratory of Mathematics, Computing and Applications, Rabat, Morocco.*
[b]*Informatics Department, Faculty of Sciences and Technologies of Mohammedia (FSTM), Hassan II University, Casablanca, Morocco*

Abstract

In this paper, we present a survey of modeling and simulation approaches to describe information retrieval basics. We investigate its methods, its challenges, its models, its components and its applications. Our contribution is twofold: on the one hand, reviewing the literature on discovery some search techniques that help to get pertinent results and reach an effective search, and on the other hand, discussing the different research perspectives for study and compare more techniques used in information retrieval. This paper will also shedding the light on some of the famous AI applications in the legal field.

Keywords: Information Retrieval; Intelligent Search; IR models; Data Mining; Natural Language Processing

1. Introduction

The expansion of data has been seen to have expanded dramatically in recent years, posing additional challenges for researchers to find innovative approaches to extract pertinent information in less time. Different techniques are employed to retrieve important data from the repository, which typically contains different types of data such as structured, semi-structured, unstructured and heterogeneous data. Various solutions, ranging from storage to information retrieval , are being studied to accommodate the fast speed of data. Despite massive developments in search understanding and technology over the last 40 years [1] , this description remains accurate and valid. the term "information" is fairly broad. Information retrieval encompasses work on a wide range of information categories and a variety of research-related applications. According to Gerard Salton, who's a pioneer in information retrieval and one of the major leaders from the 1960s through the 1990s, he suggested a definition for IR in his classic textbook [1] which said that «The structure, analysis, storage, search, and retrieval of information are all aspects of information retrieval system». The question that suggests itself is "Does the information retrieval system really retrieve a relevant document (result)?". This paper is divided into three different sections. The first section gives a brief overview of the information retrieval system. The second section describes the information search process, it presents all the phases of text pro-

* Corresponding author. Tel.: +212641513301 ; fax: +0-000-000-0000.
E-mail address: sara.ibrihich@gmail.com

1877-0509
10.1016/j.procs.2022.03.106

cessing , the methods for processing queries, ranking algorithms, and the retrieval models. The last section presents the existing related works.

2. Background

2.1. Information Retrieval Notion

Information retrieval is a pretty basic idea that almost everyone is familiar with. In today's world, the situation of a user having a need for information, translating that need into a search phrase, and then executing that search to find the information has become commonplace. Google is the most well-known example of an information retrieval system that everyone has used it. Anyone who has used such systems knows how frustrating it can be to look for specific information. Given the vast amount of thought that has gone into the design and evolution of "Google" and other search systems [2]. In other terms, we can say that Information retrieval (IR) is the procedure of representing, storing, and searching a collection of big data for the goal of extracting knowledge and access to find relevant results that satisfy the user's needs as a reaction to a user's query.

2.2. Information Retrieval Components

Figure 1 presents a basic web IR system's components.

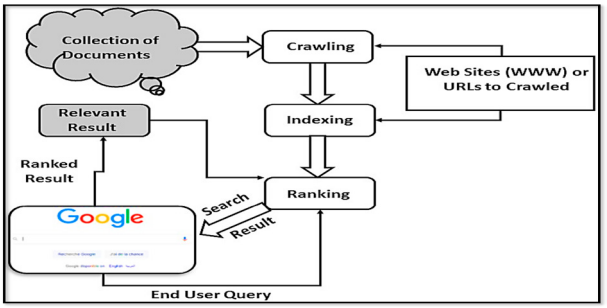

Fig. 1. Information retrieval components

As highlighted figure 1, the first components is Crawling. The crawler component is in charge of seeking and retrieving documents for the search engine. Crawlers come in a variety of shapes and sizes, but the most prevalent is the basic web crawler, which is meant to follow links on online pages in order to find and download new pages. A web crawler can be limited to a particular site, like a college, as a starting point for site search [1]. The technique of representing documents is the second component which is commonly referred to as indexing. Basically, it means that the system creates a document index. As a result, we find the query representation process. In this phase, the user writes a query in order to retrieve relevant information. After that, the system searches the index for documents that are relevant and pertinent to the query and presents them to the user, and that what's we call ranking. The last step is where the users can provide the search engine with relevant feedback [3].

2.3. Information Retrieval Challenges

The mismatch between how a user conveys the information they are seeking for and how the author of the item expresses the information he is delivering is the main challenge in information retrieval. In other words, the problem is a mismatch between the user's vocabulary (language) and the author's vocabulary (language). Besides, there are barriers to specifying the information a user requires due to limitations in the user's capability to explain what information is required. Uncertainties and ambiguities in languages are also one of the challenges that a user can face [2].

2.4. Information retrieval applications

Information retrieval (IR) systems were first created to aid in the management of massive amounts of data. Nowadays, IR systems are used by many universities, businesses, and public libraries to enable access to books, documents, journals, and other forms of data. Today, information retrieval is employed in a variety of applications [4]. The following are some examples of IR system applications:

- Media Search : In media search, we can find different types of searches such as image retrieval, blog search, news search, speech retrieval, music retrieval and video retrieval [4].

- Search Engines : For large-scale text collections, a search engine is one of the most practical implementations of information retrieval techniques. The most well-known examples are web search engines, but there are also federated search, desktop search, mobile search, enterprise search, web search and social search [4].
- Digital Library : A digital library is one where collections are stored digitally and accessed by computer. Digital content can be saved locally or remotely accessed through computer networks. A digital library is a sort of IR system [4].
- Particular field applications of information retrieval : We can find different domains that use information retrieval systems like Genomic information retrieval, IR in software engineering, Geographic IR, vertical search and of course legal information retrieval [4].

3. Processing Text

The preprocessing procedure is extremely significant and crucial. It's the initial phase in the process of text mining. In this section, we are going to discuss all the different phases that a text or document goes through.

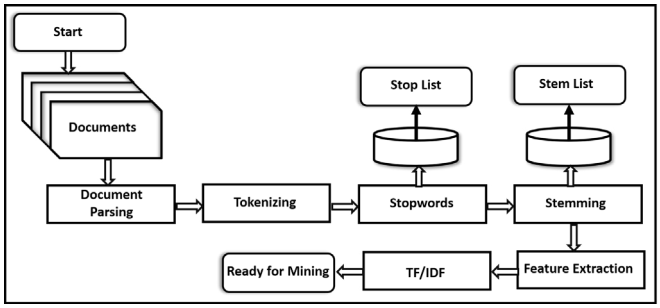

Fig. 2. Preprocessing techniques

Figure 2 highlights the various stages of text processing which are usually fulfilled by an IR engine [5].

3.1. Documents parsing

Document parsing is the process of identifying the content and form of text documents that are written and presented in a variety of languages, character sets, and forms. Often, a single document will contain many languages or formats, such as an email written in English with Spanish PDF attachments. Document parsing entails with identifying and splitting the document structure into distinct components in order to form it into unitary documents. For instance, e-mails containing attachments are separated into one document representing the e-mail and as many documents as there are attachments [5].

3.2. Tokenizing

Tokenization or lexical analysis is the operation of creating words from a series of letters (characters) in a document [1]. Normally, after parsing, lexical analysis breaks down or tokenizes the document that considered as an input stream into words, phrases, or symbols [5]. One of the problems of tokenization is the conversion of acronyms and abbreviations into a standardized format. The difficulty of tokenization varies depending on the language. Because most words in languages like English and French are separated by white space, they are referred to be "space-bound" languages. On the other hand, there are words in non-segmented languages like Chinese and Thai that not have obvious boundaries, they are referred to as non-segmented languages [6].

3.3. Stop words

The aim of this phase is to remove all the useless and insignificant common terms from the tokens streams such as articles, prepositions etc. Here are some examples of stop words, "to", "in", "at", "a" and "the". Let's take this sentence as an instance to understand how the works go, "The most apparent information retrieval applications are search engines". So, if we remove the classic stop words, we will get the following sentence: "most apparent information retrieval applications search engines"[1, 5, 7].

3.4. Stemming

Another word-level modification is stemming. The stemming component's (or stemmer's) function is to group words that share a common stem. The goal of this strategy is to eliminate multiple suffixes, minimize the number of words, ensure that stems are precisely matched, and save memory space and time. For instance, all terms present, presentation, presented and presenting can be stemmed to the term present [1, 7].

3.5. Feature Extraction

The technique of removing extraneous and superfluous characteristics from a dataset is known as feature extraction. When assigning text to one or more groups, accuracy is improved by using feature extraction techniques [8]. It's the most convenient method suggested for document classification. It helps the accuracy to get better, diminish dimensionality, and decrease processing time [8] The feature extraction algorithm depends on the vector space model, in which a sentence is represented as a dot in an N-dimensional space. Each dot's dimension indicates a different aspect of the text in digital form [9]. One method for extracting features is to use the Term Frequency Inverse Document Frequency (TF-IDF) weighted scheme [9]. The premise behind TF-IDF is that terms in a document can be separated into two categories: unique and non-unique words, regardless of whether the term is important to the document's topic or not [9]. Here is the mathematical formula how to calculate it [9]:

$$W_{i,j} = tf_{i,j} * \log(\frac{|N|}{df_i}) \tag{1}$$

This feature extraction method's general methodology is as follows. Considering a set of documents D, a term w, and a single document d ∈ D, The weight for word i in document j is $W_{i,j}$, the amount of documents in the collection is N, the term frequency of term i in document j is $df_{i,j}$, and the document frequency of word i in the corpus is df_i.

3.6. Information Retrieval models

Over the years, several retrieval models have been presented. The classical models which are the boolean and vector space model (VSM) [1]. The Boolean offers the precise match, use the logic operators (or,and,not) and it's based on a traditional set theory,which means when a word is exists in a document, its value is true or 1; otherwise, it's false or 0 [10].The VSM is applied to compute the similarity distance among document and query that are represented as vectors, by using different measures like "Cosine similarity" ,"Jaccard measure","scalar product" and "Measurement Dice" [11]. We can also find the probabilistic model which depends on the probability of the pertinent of a given query in order to rank them [10, 11]. In this model, binary vectors represent documents and queries, with each vector component signifying if the content refers or word appears in the document or not [12].

3.7. Evaluation

The ability of an IR system to return pertinent documents, as well as the accuracy and precision of these retrieved documents, is commonly measured [13]. Basically,there are two different criteria for judging the quality of IR system. According to [14] , the first one is «Precision». As Equation 2 pinpoints, it's the proportion of documents returned that are actually relevant to the query.

$$Precision = \frac{|relevantdocuments \cap retrieveddocuments|}{|Retrieveddocuments|} \tag{2}$$

The second measure is what we called «Recall». As equation 3 indicates, it's the proportion of documents that are related to the query and have been found [14].

$$Recall = \frac{|relevantdocuments \cap retrieveddocuments|}{|Relevantdocuments|} \tag{3}$$

These two measures help to calculate another information retrieval metrics which is F-measure by the following equation [1] :

$$F - measure = \frac{2 * precision * recall}{precision + recall} \tag{4}$$

4. Related work

The majority of the research has focused on statistical methods, which will be reviewed first [15]. One of its approaches is the vector space method. The similarity between documents or keywords is quantified in the field of information retrieval by appointing them into a vector of word frequencies in a vector space and determining the angle between the two vectors [16]. Sung-Hyuk Cha [17] discusses the several types of similarity metrics accessible in terms of semantic and syntactic links that are used in different information retrieval issues. Sung-Hyuk Cha et al. [18] go on to explain how vector similarities can be achieved by strengthening binary features on vectors. Vector similarity metrics were used by Jayawardana et al. to derive representative vectors for ontology classes [19]. The distance between two vectors in space determines the word similarity. There is a method suggested by Wu and Palmer

for calculating the similarity between two words in the 0 to 1 range [20], in which the cosine distance [21] was utilized as a measure of similarity. Query expansion is also one of the methods that have been studied to improve the search system and have a quick and pertinent outcome. Many crucial phrases may be missing from a query, causing a search engine or information retrieval system to respond badly or ineffectively, leading to results that are less relevant to the query. This problem was initially solved by Rocchio [22], who later presented a relevant feedback system that automatically expands original searches based on user feedback or query restructuring by providing more terms such as synonyms, plurals, modifiers, and so on [23, 24]. Roy et al. [25] suggested a word embedding techniques which are frequently utilized. It's an approach based on the distributed neural language model word2vec [26]. It used the K-nearest neighbor technique KNN [27],a non-parametric approach, to retrieve related terms to a query based on the framework. The Bayes theorem, which was stated by Thomas Bayes between 1701 and 1761 [28, 29], is the theoretical foundation of the Naïve Bayes classifier technique [30]. This latter is a generative model, which is the most used text categorization method. Table1 compare some methods used in information retrieval [30]:

Table 1. Comparison of some research work on information retrieval.

Approach	Author	Feature Extraction	Corpus	Validation measure	Limitation
Rocchio Algorithm	B.J. Sowmya et al [31]	TF-IDF	Wikipedia	F1-Macro	Only retrieves a few pertinent documents from hierarchical data sets
SVM and KNN	K. Chen et al. [32]	TF-IDF	20 Newsgroups and Reuters-21578	F1-Macro	Polysemy is not captured, and semantics and sentatics are still unsolved
Naïve Bayes	Kim, S.B et al. [33]	Weights words	Reuters-21578	F1-Macro	This strategy relies on a strong presumption about the data distribution's form

5. Legal Information System

Artificial intelligence companies are still working on ways to improve technology that can handle time-consuming activities in a variety of industries with greater speed and precision. In the legal industry, AI has already proven to be beneficial to both lawyers and clients. AI in legal practice is now evolving in a variety of applications. For instance, Canada includes new applications in the legal field such as Alexsei [1] , Blue J Legal [2], and Knomos [3] [34].In the United States, they're using AI applications like CaseText, FastCase, Judicata and LexisNexis [4] etc [34]. India is also using legal research applications such as CaseMine and MikeLegal [34].

6. Conclusion

In this paper, we gave the complete information about information retrieval. We discussed the different phases of pre-processing techniques that a document goes through which helps to retrieve efficient results. We also compared some of the existed approaches such as KNN, SVM, Naïve Bayes, and Rocchio Algorithm. In future work, we will study and compare more techniques used in information retrieval.

Acknowledgements

This work was supported by the Ministry of Higher Education, Scientific Research and Innovation, the Digital Development Agency (DDA) and the CNRST of Morocco (Alkhawarizmi/2020/25).

References

[1] W. Bruce Croft, Donald Metzler and Trevor Strohman.(2015) "Search Engines Information Retrieval in Practice", Pearson Education, Inc.
[2] G. Kowalski.(2011) "Information Retrieval Architecture and Algorithms", Springer (eds) , Boston, MA.
[3] Lal, N., Qamar, S., and Shiwani, S.(2016) "Information retrieval system and challenges with dataspace." *International Journal of Computer Applications* **147** (8).
[4] Roshdi, A., and Roohparvar, A.(2015) "Information retrieval techniques and applications." *International Journal of Computer Networks and Communications Security* **3** (9): 373–377.

[1] http://www.alexsei.com/
[2] www.bluejlegal.com/ca
[3] http://www.knomos.ca/
[4] https://www.lexisnexis.com/

[5] Ceri, S., Bozzon, A., Brambilla, M., Della Valle, E., Fraternali, P., and Quarteroni, S.(2013) "The information retrieval process" *In Web Information Retrieval* : 13-26.
[6] Kannan, S., Gurusamy, V., Vijayarani, S., Ilamathi, J., Nithya, M., Kannan, S., and Gurusamy, V.(2014) "Preprocessing techniques for text mining." *International Journal of Computer Science and Communication Networks* **5** (1): 7-16.
[7] Vijayarani, S., Ilamathi, M. J., and Nithya, M.(2014) "Preprocessing techniques for text mining-an overview." *International Journal of Computer Science and Communication Networks* **5** (1): 7-16.
[8] S.Vidhya , D.Asir Antony Gnana Singh and E.Jebamalar Leavline.(2015) "Feature extraction for document classification." *International Journal of Innovative Research in Science,Engineering and Technology* **4** (6): 50-56.
[9] Dzisevič, R., and Šešok, D.(2019) "Text classification using different feature extraction approaches." *Open Conference of Electrical, Electronic and Information Sciences* : 1-4.
[10] Gudivada, V. N., Rao, D. L., and Gudivada, A. R (2018) "Information retrieval: concepts, models, and systems", *In Handbook of statistics Elsevier* **38** : 331-401.
[11] Nadia, L.(2014) "Design and implementation of information retrieval system based ontology." *In International Conference on Multimedia Computing and Systems (ICMCS)* : 500-505.
[12] Sharma, M., and Patel, R.(2013) "A survey on information retrieval models, techniques and applications." *International Journal of Emerging Technology and Advanced Engineering* **3** (11): 542-545.
[13] Bassil, Y.(2012) "A survey on information retrieval, text categorization, and web crawling." *arXiv preprint arXiv* **1** (6): 1-11.
[14] Saini, B., Singh, V., and Kumar, S.(2014) "Information retrieval models and searching methodologies: Survey." *Information Retrieval* **1**(2):20.
[15] Terra, E. L., and Clarke, C. L.(2003) "Frequency estimates for statistical word similarity measures." *In Proceedings of the human language technology conference of the North American Chapter of the Association for Computational Linguistics* : 244-251.
[16] Salton, G., and Mcgill, M. J.(1986) "Introduction to modern information retrieval."
[17] Cha, S. H.(2007) "Comprehensive survey on distance/similarity measures between probability density functions." *International Journal of Mathematical Models and Methods in Applied Sciences* **2** (2): 1.
[18] Cha, S. H., Yoon, S., and Tappert, C. C..(2005) "Enhancing binary feature vector similarity measures."
[19] Jayawardana, V., Lakmal, D., de Silva, N., Perera, A. S., Sugathadasa, K., and Ayesha, B..(2017) "Deriving a representative vector for ontology classes with instance word vector embeddings." *In Seventh International Conference on Innovative Computing Technology* : 79-84.
[20] Wu, Z., and Palmer, M.(1994) "EV Verbs semantics and lexical selection." *In Proceedings of the 32nd annual meeting on Association for Computational Linguistics* : 133-138.
[21] Qian, G., Sural, S., Gu, Y., and Pramanik, S.(2004) "Similarity between Euclidean and cosine angle distance for nearest neighbor queries." *In Proceedings of the ACM symposium on Applied computing* : 1232-1237.
[22] Robertson, S. E.(1977) "The probability ranking principle in IR." *Journal of documentation* **33** (4): 294-304.
[23] Diaz, F.(2016) "Pseudo-query reformulation." *In European Conference on Information Retrieval* : 521-532.
[24] Salton, G., and Buckley, C.(1990) "Improving retrieval performance by relevance feedback" *Journal of the American society for information science* **24** (5): 355-363.
[25] Roy, D., Paul, D., Mitra, M., and Garain, U.(2016) "Using word embeddings for automatic query expansion." *arXiv preprint arXiv:1606.07608*
[26] Azad, H. K., and Deepak, A.(2019) "Query expansion techniques for information retrieval: a survey." *Information Processing and Management***56** (5): 1698-1735.
[27] Li, L., Weinberg, C. R., Darden, T. A., and Pedersen, L. G.(2001) "Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method." *Bioinformatics* **17** (12): 1131-1142.
[28] Hill, B. M.(1968) "Posterior distribution of percentiles: Bayes'theorem for sampling from a population." *Journal of the American Statistical Association* **63** (322): 677-691.
[29] Pearson, E. S.(1925) "Bayes'theorem, examined in the light of experimental sampling." *Biometrika* **17** (3/7): 388-442.
[30] Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., and Brown, D.(2019) "Text classification algorithms: A survey." *Information* **10** (4): 150.
[31] Sowmya, B. J., and Srinivasa, K. G.(2016) "Large scale multi-label text classification of a hierarchical dataset using Rocchio algorithm." *In 2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS) IEEE* : 291-296.
[32] Chen, K., Zhang, Z., Long, J., and Zhang, H.(2016) "Turning from TF-IDF to TF-IGM for term weighting in text classification.." *Expert Systems with Applications* **66** : 245-260.
[33] Kim, S. B., Han, K. S., Rim, H. C., and Myaeng, S. H.(2006) "Some effective techniques for naive bayes text classification.." *IEEE transactions on knowledge and data engineering* **18** (11): 1457-1466.
[34] Langlois, P., and Titah, R.(2020) "Utilisation et impact des outils d'intelligence artificielle dans des contextes de cyberjustice." *Doctoral dissertation, HEC Montréal* .