ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 1267-1277 (2020)

www.astesj.com

ASTESJ

ISSN: 2415-6698

Towards a Documents Processing Tool using Traceability Information Retrieval and Content Recognition Through Machine Learning in a Big Data Context

Othmane Rahmaoui*, Kamal Souali, Mohammed Ouzzif

Computer Science Department, Hassan II University, Casablanca, 20000, Morocco

ARTICLE INFO

Article history:
Received: 18 November, 2020
Accepted: 07 December, 2020
Online: 16 December, 2020

Keywords:
Document Management
Data traceability
Information Retrieval
Machine Learning
Recommendation
Spark

ABSTRACT

In 1980, an application was developed to track, manage and store documents in electronic format. Scan technology has enabled organizations to digitize papers for easier document storage and tracking. Document management tools have since developed by introducing new functionalities, related to security, users services, workflow and audit. Our research is part of the context of improving the efficiency of document processing by proposing an approach using information traceability retrieval and content recognition techniques through machine learning. In this sense, we started by proposing the exploitation and extraction of relationships between documents based on the traceability links and the calculation of similarity using information retrieval. Then, in order to improve the processing of documents, we proposed a contribution of the use of recognition of content techniques by machine learning approaches. Thus, the visualization of the results, according to user profiles, motivated us to offer recommendations dedicated to the document management system. A Big Data environment is proposed because of the exponential growth of data and also to our needs concerning the analysis and the distributed calculation of the voluminous masses of data.

1. Introduction

The definition of a document according to ISO (International Organization for Standardization); "Is the set consisting of an information medium and the data recorded on it, in a generally permanent form and readable by humans or by a machine". A document in electronic format is a file created, modified and readable by a dedicated computer application; it can be an audio, an image or any data file. Electronic documents are the precious treasure and the important part for the reuse and preservation of valuable data in storage based on operators of compression, decompression, encryption and decryption [1], [2]. From these definitions, we notice that any document, no matter how electronic or classic, it keeps two important properties; the information which represents the content processed at the document level and the metadata part which offers the possibility of introducing the notion of object at the level of the design of a document [3]. In the Information Technology (IT) domain, traceability is becoming an obligation for companies offering applications, systems or services that meet market requirements. At the IT project level, changing needs is essential in order to develop a good product that meets all customer requirements, in which case traceability is an essential element for managing changes as well as analyzing its impacts. However, traceability must be proposed to show that a methodological development process has been respected and to justify that the developed tool is ready for use [4]. Using what we have called traceability information retrieval, for the purposes of this article, is a process of producing a list of relevant documents as a response to a user's request by comparing the request to an automatically generated index of the content of documents in the tool [5]. These documents can be consulted subsequently to be processed in the same tool. Today everyone uses tools and applications using information research techniques, for example the search engines Google, Yahoo and Bing.

Nowadays, organizations are constantly seeking to be more flexible and optimized to cope with the revolution in the digital world. Each organization aims for organizational excellence through different means to satisfy either the project teams or their customers or their suppliers [6]. To meet expectations, an effective policy and a strategic vision must be created in a clear way through the structuring of processes, documents and resources. According to management standards [7], the updating of documents plays a very important role in the continuous improvement of the

*Corresponding Author: Othmane Rahmaoui, Hassan II University, ESTC RITM Lab., ENSEM, Casablanca, Morocco, othmane.rahmaoui@gmail.com

company, because the documentation, updated, guarantees optimal efficiency unlike documentation that discards outdated information and so that will make it obsolete. Documentation management is one of the important levers for building quality management; it allows the organization of a set of information, documents and records used in projects or in the various departments of the 'business [8]. A document can be in paper or electronic format.

In order to protect and ensure the validity of easily accessible data as well as keep traceability on the operations and activities carried out, we have thought of improving the processing of documents. However, we started by researching and formalizing methods and procedures necessary and sufficient to control the information and documents useful for the entire organization (entity, service, project, etc.) as well as the technologies that can be used in this direction. The proposed approach targets important issues in order to limit the risks of loss of documents and to make work collaborative with the sharing and capitalization of resources as well as communication between employees and accessibility to information, without forgetting security and reliability as well as traceability management [9], [10]. Our approach saves time in the search for information or documents, increases efficiency, increases the satisfaction of project teams (users) as well as reduces paper volumes, while using modern and efficient techniques through which we called ; IR Traceability Recommendations: Recommendations based on data traceability and information retrieval. Our approach deals with the case of document management that concerns the business domain, as we can generalize it in order to apply it to any sector and in any organization.

For electronic document management, improving the flow of documents has been widely studied in several research works based on the integration of one or more techniques through the implementation and proposal of tools for monitoring and storage as well as document management. The document management administrator is always required to monitor and control the flow of documents processed [3], [11]. Few works offer tools for automatic and incremental improvement of electronic document management. These contain unstructured data and are difficult to operate in another system and even the creation of relationships between information from different documents is difficult to manage, which is why we sought to propose a transformation of documents into data in order to facilitate the processing and development as well as the analysis of data and therefore improve the management of documents in an automatic manner.

In recent years, Big Data with information retrieval has become a more popular area of research [12]. Big Data is a collection of heterogeneous unstructured and structured data. The heterogeneity, volume and speed at which data is generated make the processing and analysis of big data problematic [13]. The traditional database system, warehouses and analysis tools fail to process this type of data [14]. Big Data with information retrieval techniques is an emerging approach not only because of the sheer volume of data, but also because of the kind of unstructured nature. Data related to the user's request should be retrieved with information research. According to experts, employees spend around 11 hours a week dealing with document-related issues, resulting in low productivity and poor customer service. Hence we came up with the idea of harnessing the potential of machine learning, so that companies can revamp and improve the traditional document management system. Machine learning can also be a powerful data mining tool [15], [16]. Using this technology, companies can develop a way to help their employees extract relevant data, classify and process documents automatically, improve data quality and analyze data. Document management by machine learning will therefore save a lot of time and effort. Recommendation techniques also provide new opportunities for the search for personalized information on this work. The recommendation, on the other hand, also serves to reduce information overload, which is a very common problem in information retrieval tools and allows users to have access to services or products that are not readily available to users on the tool [17]-[19].

In this article, we present an improvement of a proposed document management system using traceability and information retrieval with content recognition techniques in a big data context.

Our contribution consists in the improvement of document management systems. The remainder of this article is organized as follows ; The Sect. II presents a state of the art about document management, traceability information retrieval and content recognition through machine learning research area. The literature made was chosen based on recent research and discussion of important topics related to document processing, in order to be useful in relation to the objective of our proposed approach. We present the use of traceability and information retrieval in a big data environment in Sect. III. Machine learning techniques and recommendation are described in Sect. IV. The Sect. V describes the suggested document management system and we conclude in Sect. VI.

2. State of the art

2.1. Document management system

Document management is defined as "the set of techniques for organizing, managing and distributing documentary information in electronic form" [20]. Document management is actively developing especially within companies, because it covers the administration of the document flow within all departments of the company. In the literature [11], [21], [22], we can define document management system as seamlessly automating the document lifecycle process and providing tools to securely manage a large range of documents. Document management system uses statistical algorithms to classify and search documents. Document Management is actively developing especially within companies, because it covers the administration of the document flow within all departments of the company. The objective of this work is to propose techniques to improve the efficiency of document management on the one hand and to minimize the total cost of selecting and recommending documents on the other hand. For document management, improving the flow of documents has been widely studied in several research works based on the integration of one or more techniques through the implementation and proposal of tools for monitoring and storage as well as document management [8], [1].

2.2. Information retrieval based on traceability

In the world of computer science research, traceability is a modern and very recent term, in dictionaries this term appeared

only around the year 2000 while traceability has been known for a long time, its definition differs according to the fields and sectors (Industry, Food, Production, Logistics ...) [23]. The ISO-9001 standard of the international organization for standardization, defines traceability as: 'the ability to trace the history, application, use and location of a product or its characteristics using data from 'recorded identification', Indeed, this definition generates two important phases: the first concerns the identification of the item by a notation and the second concerns the existence of a recording of the data relating to this item on a support which has itself a traceability. The literature shows that there are problems related to the definition and understanding of traceability. the definition of which changes depending on the way it has been proposed. According to the dictionary, traceability is the "possibility of identifying the origin and reconstructing the journey of a product, from its production to its distribution". Researchers say that traceability is not a new concept but a practice that we must implement in order to respect the norms and rules of law [24]. However, in recent years, it has become a necessity in areas where the safety and security of consumers have challenged, particularly in the medical and food sectors [25]-[27]. Traceability is not a new concept but it represents the necessity of results which imposes the replacement of papers and pens by means and techniques more reactive and especially more efficient, traceability comprises two essential elements; tracing and tracking; The purpose of traceability is to provide proof at all times of the conformity of an element and its components as well as of its origin. Tracking represents the way to be able to save and know the paths of elements at the level of the data flow and the tracing is the fact of identifying the origin of a specific data at the level of the flow.

Information retrieval is a field allowing the search for relevant information from any type of data (audio, video, document, article or image). Traditional methods of finding information involve breaking down data into subsets or clusters across dimensions depending on the given problem. The Information Retrieval (IR) approach reduces the effort and time prescribed to create traceability links between artifacts [28] [29]. A search engine is an example of using the IR approach; it can search billions of web documents to answer a very specific query. Information retrieval techniques make it possible to solve several problems related to the processing carried out in a particular domain as well as to its optimization [30]. We have, in our case study, a large remarkable mass of documents and objects, and therefore to find an element adaptable to a need for information expressed by a user using a search query, we have thought of using the algorithms of the traceability information retrieval with recovery of traceability links.

2.3. Machine learning and recommendation

Machine learning has become one of the key words in recent research, machine learning represents an influential source of applications to automate the tasks of information retrieval. the latter offers several features that would fit into machine learning, but really what drives applications of machine learning to IR is not so much developments in machine learning technology as changes in our work and environment which requires new modes of operation [31]. Machine learning is an important discipline due to the digital transformation which has led to the production of large volumes of data of different types and formats. Machine learning can be seen as a branch of artificial intelligence. Indeed, a system which finds it difficult to learn can hardly be considered intelligent. The ability to learn from experiences is essential in a system designed to adapt to an environment that changes often. Artificial intelligence, defined as 'the set of techniques used to build machines capable of demonstrating behavior that can be described as intelligent' [32], also calls on cognitive sciences, to logic, to engineering to electronics, and much more. Artificial intelligence measurements are used by different tools and applications to collect and process and also to communicate and share useful information from datasets. Arthur Samuel defined machine learning, in 1959, as a "field of study that gives computers the ability to learn without being explicitly programmed" [33]. Machine learning (ML) is a concept that allows a computer program to be learned as well as to adapt to new information and new data without any human intervention. With the gradual use of technologies like cloud computing and advanced storage capabilities, the huge amount of data becomes easily accessible and available. Machine learning is the discipline of the computer science that concerns the analysis and interpretation of models and data to enable learning and processing as well as reasoning and decision making without thinking of human intervention. In a simple way, machine learning allows the user to feed an immense amount of data into an algorithm and ask the machine to analyze and make decisions as well as recommendations based on the entered data [34]. If corrections are identified, the algorithm can incorporate this information to improve future decision making. Decisions about data are increasingly making the difference between staying up to the competition and falling behind. Machine learning can be the key to unlocking the value of business data and customer and make decisions that allow a company ahead of the competition.

2.4. Literature and related work

Several works talk about traceability or information retrieval and also about content recognition through machine learning in a big data context, in literature we are finding :

About documents classification in [35], author proposed a review about a wide variety of methods on document classification task for German text with an evaluation of different approaches; Perceptrons, Discrete Naïve-Bayes, MC4, 3 Nearest-Neighbors, Racchio Centroid and Support Vector Machines. In [36], the author have proposed a framework for modeling link distributions by using a structured logistic regression model capturing both content and links in order to improve text classification. In [37], the author have presented TRAIL (TRAceability lInk cLassifier) as an approach for maintaining traceability information (traceability links between pairs of artifacts) to train a machine learning classifier which is then able to classify the links between any new or existing pair of artifacts as valid or invalid (two artifacts are related or no). So, document classification is an example of machine learning in the form of natural language processing (NLP), in the real world numerous more complex algorithms exist for classification such as support vector machines (SVM), Naïve Bayes and Decision Trees. There are also works that use information retrieval with several traceability methods; in [38], the author have presented a study to analyze the equivalence of information retrieval methods for automated traceability link recovery; Jensen-Shannon (JS), Vector Space Model (VSM),

Latent Semantic Indexing (LSI) and Latent Dirichlet Allocation (LDA). The result show that while JS, VSM and LSI are almost equivalent. In [39], the author have proposed an approach for improving requirements tracing based on framing it as an information retrieval (IR) problem, they have started with a classical vector space model algorithm but they found that this algorithm does not outperform analysts or existing tools in terms of recall or precision. In [40], the researchers describe a new technique called DROPT (Document Ranking Optimization) developed for documents retrieved from a corpus with respect to document index keywords and the query vectors, the purpose of DROPT id to reflect how human users can judge the context changes in IR result rankings according to information relevance. In [41], the author represent an algorithm based on classical mathematical expressions for calculating similarity between groups, known as the cosine, jaccard and dice formulas. In [42], an application of information retrieval to software maintenance and in particular to the problem of recovering traceability links between the source code of a system and its free text documentation by using an IR method (Vector Space), and in [43] a source code parser has been described to create a syntax tree as a processing step for information retrieval to recover traceability links between source code and documents. Several works of recovering of traceability links between documentation and source code by using Latent Semantic Indexing (LSI) like an information retrieval technique.

About the use of Machine learning, in [44] the author have proposed a predictive model for classifying potential traceability links in a system as either valid or invalid, by using a Machine Learning approach and features such as text retrieval (IR), rankings and query quality (QQ) metrics. In [45], the author also uses a machine learning approach for generating requirements traceability relations, this approach is based on a learning algorithm that produces traceability rules which are able to capture link between requirements statements specified in natural language and object models.

3. Improving document management using traceability and information retrieval in a big data environment

3.1. Context

The main goal of this work is to build a scalable model for classifying documents and extracting information from them. In a corpus of documents, the model extracts the text from each document and applies vectorization to it. Vectorization consists of counting the number of occurrences of each word appearing in a document. The vectorized representation is then entered into the model for prediction. Vectorization includes several algorithms and techniques in order to extract as much information as possible. Once this information is extracted and after dividing the data into a test set and a validation set, we apply different classification models on the data (VMS, KNN and decision trees). With arbitrary tests, the measurements showed very attractive results: precision, recall and F-measure between models, but the best result was recorded using TF / IDF with an accuracy of 0.934, an accuracy of 0.942, a recall of 0.905, and an F measure of 0.921.

When a user initiates a query, the query has multiple terms, and some of those terms may be more important than others. TF-IDF is a popular technique for measuring and getting an idea of the importance of a term in relation to a document. The TF is used to measure how often a word appears in a document. The IDF measures the importance of a word (Table 1). The TF-IDF value increases if a term appears in the document a number of times while taking into account the number of existing documents in the corpus that contain the same term.

Table 1: Example of frequency of appearance of a term in a document

	Document 1	Document 2	Document 3
Term 1	5	2	1
Term 2	1	0	0
Term 3	3	1	0
Term 4	0	4	3

The following formula will allow us to calculate the weights of the terms in each document:

$$\mathrm{Poid}(t_i, d_j) = \mathrm{TF} * \mathrm{IDF} \quad (1)$$

with:

$$\mathrm{TF} = \frac{f(t_i,d_j)}{N} \quad (2)$$

$f(t_i,d_j)$ is the number of occurrences of the term t_i in the document d_j and N is the total number of terms in the document d_j.

and

$$\mathrm{IDF} = \frac{\log\left(f(t_i, d_j)\right)}{M} \quad (3)$$

$f(t_i,d_j)$ is the number of occurrences of the term t_i in the document d_j and M is the total number of documents in the corpus.

We say that two objects are similar if their vectors are confused, so the similarity used in our work is the cosine similarity, this measure uses the complete vector representation based on the frequency of objects or words. The formula is defined by the ratio between the scalar product of the X and Y vectors and the product of the norm of X and Y:

$$\mathrm{Sim}_{\cos}(X, Y) = \frac{\sum_{i=1}^{n} x.y}{\sqrt{(\sum_{i=1}^{n} x^2)}.\sqrt{(\sum_{i=1}^{n} y^2)}} \quad (4)$$

A document is made up of a joint agreement of terms that have various patterns of occurrence. Text classification is an important task especially for document processing, but with the expenential growth of data, there are algorithms that can improve classification efficiency while ensuring the high precision desired. The purpose of classification is to organize documents by sets of documents that are similar, and theoretically the documents are said to be close. This phase aims to classify documents for the sake of research and extraction of useful information in an efficient manner. In this phase we propose to use VMS with the KNN (K Nearest Neighbor) algorithm and the decision tree algorithm in parallel, which makes it possible to put a document in a dominant set among its k nearest neighbors. A similarity measure makes it possible to determine the documents closest to the request launched by the user or, in the

case of classification, to determine the sets closest to the document to be classified.

KNN is a popular algorithm in the field of classification. Its use as well as its results place it among the best classification techniques in [46] [47]. As for the decision tree, it is an algorithm that makes it possible to perform an iterative partitioning of the entire feature space; The tree predicts the same label for each lower category and each category is chosen by selecting the best division from a set of possible divisions, which maximizes the information gain at a tree node [48]. However, a decision tree in the sense of our work will allow to classify data or observations already labeled as well as the visualization of the processed data; The tree starts with a source, where there are all the observations, then several branches in series show the nodes and at the end we find the leaves which represent the classes to be predicted. The major goal of a decision tree is to produce user-understandable classifications.

When we have a request, we must find the relevant objects, however the way to evaluate a document, whether relevant or not, is to calculate the similarity between the request and the object. Before calculating the similarity, it is important to index, clean and format all the contents of all the documents by using the techniques of the recovery of the traceability of the data (links, who, when, how, where ... etc.) as well as the classification using a machine learning. The vector representation will also allow us to facilitate the use of the requests made where each element of the vector represents the weight of each term or concept in the document or in the query. The objective is to extract all the terms (t) or concepts from our corpus and for each document (d) we build a vector (v) which represents it, if a term exists in the document we calculate its weight, otherwise we put 0, at the end of the operation we will have a vector for each document to calculate the similarity between the documentation and the purpose of the research done.

The traceability processed in this work represents an entire history of the operations carried out on the documents. Once an operation is performed by a user, this action is tracked (consultation, modification, deletion…). The traced data makes it possible to know, at any given moment, the life cycle of the documents as well as the associated operations. Each document with unstructured data must be converted into a format that can be analyzed in order to structure it and extract useful information. The content of a document can be represented as a set of terms such as words, sentences or other entities, each term will have its weight, which gives its utility in relation to this document. The main objective of document exploration is then to allow administrators to extract information from textual resources and to manage operations such as retrieval of traceability links and classification.

3.2. Traceability Information retrieval

In terms of our approach, we take into consideration any operation, any action and any event such as a trace stored in the database. We present a traceability link by a triple (source document DS, target document DC and similarity S). In order to apply IR techniques for the proposed approach, we consider all artifacts as a text document first. The IR allowed extracting all terms from a DS source document by calculating the similarity to another DC target document, depending on the document type and after processing, in order to extract the information. The high similarity between two documents suggests a potential link between them. For each DS, preprocessing done according to the type of document (image, video, text, etc.) for which we have thought of using machine learning techniques, we will generate the necessary traceability information in the database.

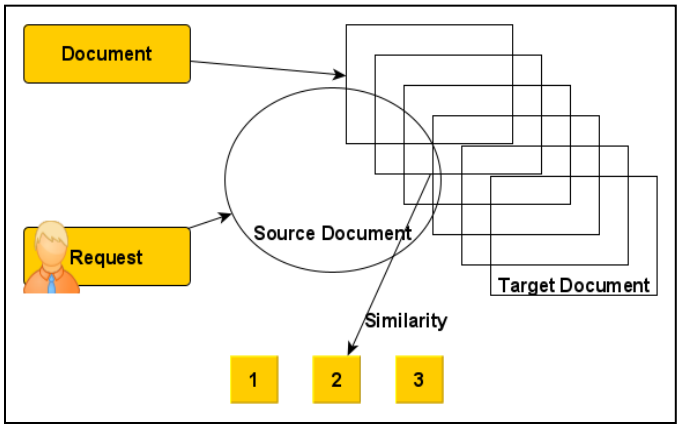

Figure 1: Similarity calculation and recovery of traceability links

We take as input, the construction of a matrix which concerns the terms per document (m × n), with m the number of all the terms appearing in the documents and n the number of documents. For each cell c_{ij} of the matrix, with i represents the term number and j the document number. We calculate the weight of each term and by applying the IR we calculate the similarity between two documents. The traceability links are created for each two documents with the value of their similarity. The latter is calculated by the positive cosine. For the recovery of traceability links, we aim to trace everything related to documents, even the traceability links in the document itself if it is divided into several parts.

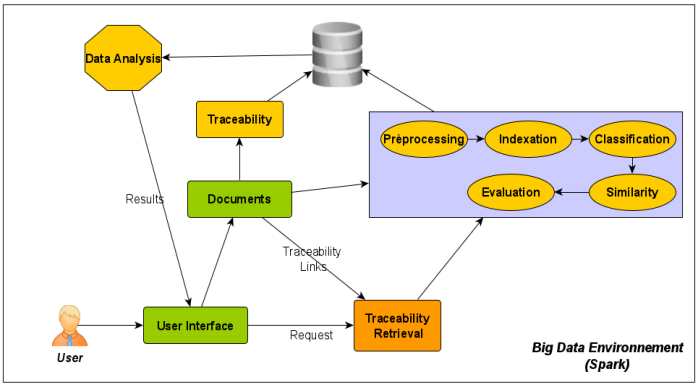

Figure 2: Process of the proposed approach

The interest in the need in Big Data is increasing day by day. In the age of big data, variety dominates volume and speed. Several difficulties and challenges may arise in acquiring, storing, processing and retrieving data. Companies like Google, Amazon, Facebook and YouTube have an advantage over others in personal development and competition due to their large amount of data. It should be remembered that processing data and finding information is as important as storing data. At the enterprise level, information can be structured or unstructured. In the first case, the information is presented in the form of a database and in the second case; the information is presented in the form of documents. Documents capable of keeping secrets that sometimes occupy special confidentiality in the company.

The architecture of the proposed approach is represented as shown in Figure 2.

The traceability of operations as well as the recovery of traceability links between documents is an important phase in order to reduce the number of documents processed when the user launches a search query. All of the document techniques algorithms help automate the process to have better results and identify relevant documents with very good precision. The Spark environment takes into account the sheer mass of data and its exponential growth, as well as the analysis done, in order to meet needs and help make decisions.

For the implementation of our approach we have carried out tests on tools to show what the approach can guarantee for the improvement of document management. We used KNIME (KoNstanz Information MinEr) [49], a software developed in 2004 by the German University of the city of Constance (Konstanz). It is open source software written in java. KNIME is a tool that allows the design and execution of workflows using nodes that are already predefined as components of the platform. What is also huge is that this platform also allows the processing of large volumes of data depending on the resources of the user's machine (storage capacity, RAM, etc.); That is, there are no restrictions with regard to the volume of data or with regard to the platform itself. KNIME is an easy-to-use application, what you just have to respect is the order of operations as well as the logic of the process so you have to choose the node corresponding to the task seeking to be executed in order to make the execution of the workflow successful ; In KNIME, a node represents an entity that performs a well-defined task.

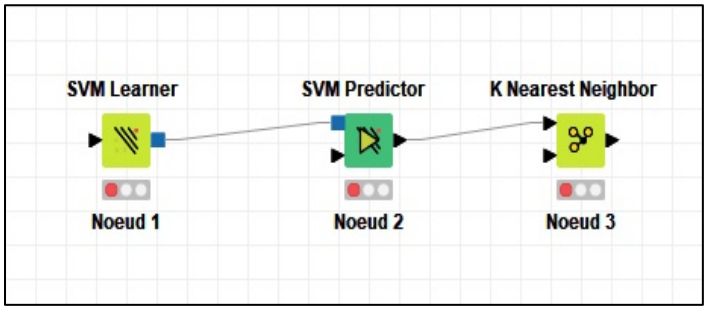

Figure 3: Example of node in KNIME

Figure 3 shows an example of nodes corresponding to the use of the VSM model with the KNN algorithm; The first node is used to train an SVM model against the input data. The second node uses the SVM model generated by the first SVM node to predict the output for given values. The third node classifies a dataset based on the KNN algorithm using the training data. A node consists of input, output and its state; if the state is in red it means that the node is not configured, if the state is in yellow it means that the node is well configured and if it is green, it means that the node is well executed, otherwise there is an error.

The visualization of the precision statistics shows in figure 4, the good precision taken with respect to the sets processed using the three types of VSM algorithm, Decision tree and KNN.

This approach is presented in the form of examples of tests in order to show the interest and the advantages brought to improve the management of documents by a system which allows to have a global vision on the history of the operations made on the documents. , as well as the relationship between them. On the other hand, it also makes it possible to facilitate the processing and application of information retrieval algorithms in order to extract adequate information according to specific needs. The use of the big data environment, especially Apache Spark, will allow to perform distributed calculations on a large volume of data and in a reasonable time. Our approach does not differ much from other approaches, but the proposed process has never been discussed in relation to electronic document management systems. On the one hand, we offer an important management of the traces made on the documents as well as the links between them, and on the other hand, by increasing the speed and the ease of treatment we offer the techniques of information retrieval as well as analyzes the distributed computing offered by Spark. While, the other works generally offer a classic system for managing documents manually. The works that offer tools for the electronic management of semi-automatic documents are rare, and those already existing are limited to private and commercial use.

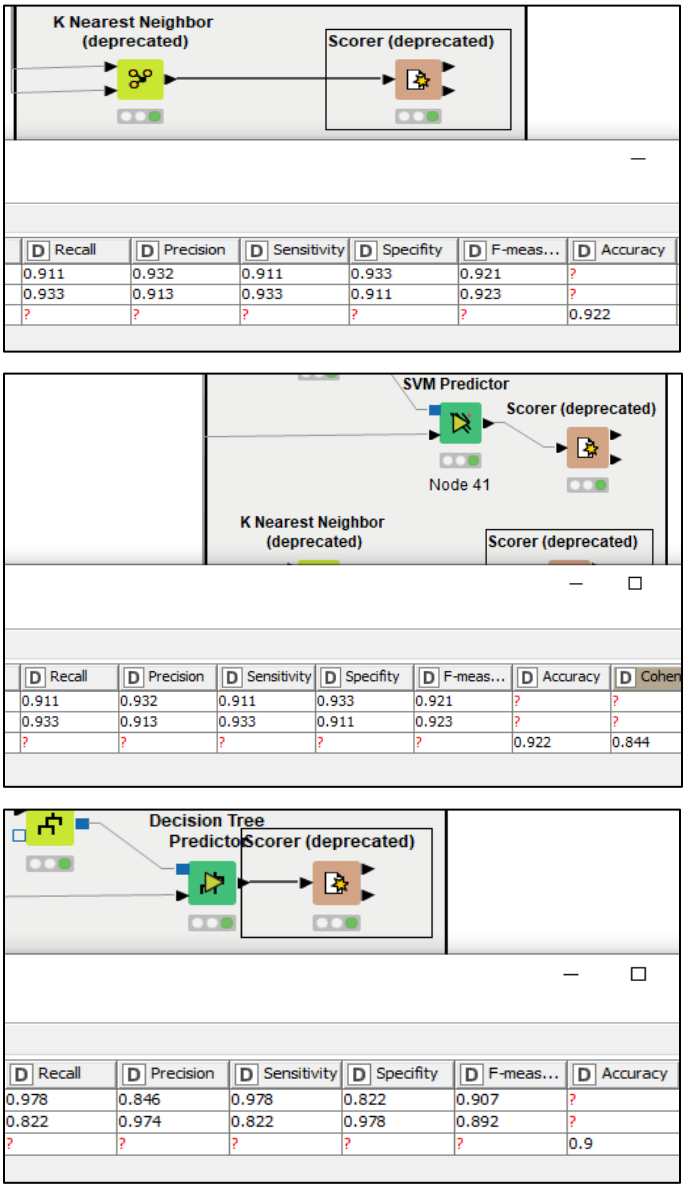

Figure 4: Precision statistics

4. Using Contents Recognition techniques in document management

4.1. Context

Several approaches to machine learning have been applied successfully to real world problems. Indeed, the choice of method in each case seems to be based largely on the experience and preference of the researchers involved. This success is also apparent in information retrieval applications, the different paradigms being used in similar proportions to those in other fields. Much of the information retrieval work can be automated. Processes such as document indexing and query refinement are usually done automatically, while document classification and term selection are more often done semi-automatically or manually. However, algorithms that classify documents by types to index information and model user interests to help them formulate queries, reduce workload, and ensure more consistent behavior. These algorithms are based on machine learning, a dynamic and growing field of computing.

The objective of this contribution is to propose an approach to improve document management based on information traceability retrieval but also on effective machine learning to extract data, create a model and provide a classification. Implementing such a solution could reduce administration costs, speed up the document delivery process and improve customer satisfaction in businesses. At the level of our proposal, we present a kind of multimedia management system; the use of machine learning also helps us in the automatic classification and extraction of data when a user adds any document (image, video, audio or text) to the corpus, using classification algorithms and content recognition algorithms. Until now, most companies simply scan these documents, index them with a date and number, and store them in a repository. There is currently a wide range of specialized algorithms for certain parts of document analysis. In large scale applications, these approaches must cope with the wide variety of documents. Machine learning algorithms are a possible treatment in this situation. From a set of documents, these algorithms are able to extract a lot of relevant data, and even they are able to weight this data in such a way that changing some information does not lead to a loss of performance. We aim in this work to propose an automatic classification of documents in predefined categories as well as by document type (image, video, audio or text). Supervised learning techniques are used for automatic classification, where predefined category labels are assigned to documents based on the probability suggested by a set of document training. Some of these techniques are described below. Several algorithms or combination of algorithms as hybrid approaches have been proposed for the automatic classification of documents. Among these algorithms there are VSM, decision trees and KNN (k-nearest neighbors) and their hybrid system with the combination of different other algorithms and techniques [50].

Computer systems that could learn to predict from a set of data and improve without having to be reprogrammed were a dream until recent years. But now, it has been made possible through machine learning. Today, machine learning is the most widely used branch of artificial intelligence which is adopted by major industries to benefit their businesses. Machine learning unveils a breakthrough science in how computers can learn and make predictions. It has applications in various industries and it is widely used. Machine learning has grown in popularity since its inception and it won't stop immediately. Data scanning is on the increase to help cut costs and save time, but simply converting documents using a scanner into an image or PDF file is not enough. A document in printed or handwritten form can be modified or edited at any time, which is not possible for a converted image. Thus, content recognition becomes a requirement to add search and edit functionality in a scanned document.

The most important documents and files today are stored in digital format. This is because digital storage is simple and does not involve physical storage space. In addition, the digital retrieval of documents is much easier than with filed paper documents. However, scanned documents are treated as image files and cannot be searched or edited because they are not machine readable. Optical Character Recognition (OCR) is a solution to convert these scanned documents into machine-readable text files. Files processed with OCR can be searched and edited as needed. Recognizing content in a document management system will therefore be a new approach that promises to transform the way businesses manage document processing. It will aim to recognize invoices, tax forms, survey forms and various other business and administrative documents which may be formal or loosely structured with proper storage and retrieval of these documents for business purposes.

4.2. Machine learning for document management

The essential goal of using machine learning in this work is to apply it to multimedia documents and also to have a classification, initially, by type of document (Image, video, audio or text), then and thanks to these content recognition algorithms we will try to extract the information that will be used on the information retrieval part. After this processing on the content of the documents, an automatic categorization is proposed. The automatic classification of documents is quite a complicated task. All the varied documents have different visual parameters: size, format, shade of paper, font size and type. The fact that some of the documents are handwritten complicates things and it was very difficult to distinguish the handwriting. Documents vary to a great extent and new documents are expected. This is why it is almost impossible to solve this task by software implementation of algorithmic solutions but by offering machine learning, this processing by learning algorithms seems reasonable.

Recognizing content in a document management system will therefore be a new approach that promises to transform the way businesses manage document processing. It will aim to recognize invoices, tax forms, survey forms and various other business and administrative documents which may be formal or loosely structured with proper storage and retrieval of these documents for business purposes.

4.3. Recommended documents

Spark MLlib is an Apache Spark module that provides machine learning primitives as APIs. Machine learning typically processes a large amount of data. Spark's basic IT framework is a huge plus. Additionally, MLlib provides most of the popular machine learning and statistics algorithms. This greatly simplifies the task of working on a large-scale machine learning project. MLlib offers recommendation techniques for unsupervised learning and

especially in the part of clustering that will be proposed in this work. Recommendation systems are software tools and techniques that provide suggestions of documents for a user to use. For our recommendation tool to be effective, it must have a comprehensive and properly cataloged resource repository. In addition, it must be precise to understand the user's needs and must essentially profile the user correctly. Ideally, you should take into account that a user's preferences are constantly changing. To address this need, we provide an interactive document recommendation tool that observes the different themes that arise in user work over time and continues to learn user preferences continuously based on feedback. of the user on the documents recommended to him. We then estimate the similarity between the query and each document using processed by the IR process presented in the first contribution. To take into account user preferences, we keep track of user documents to ensure traceability throughout the process.

Our recommendation phase depends on the user's profile. A profile can be made up of several types of information. In this proposition, we distinguish two types of information:

- A template of user settings, that is, a description of the types of documents that are of interest to the user. There are many other possible representations of this description, but a common representation is a function that, for any item, predicts the likelihood that the user will be interested in that document. For efficiency reasons, this function can be used to retrieve the n documents most likely to interest the user.
- A history of user operations with the document management system. This may include documents the user has viewed. The history also includes the recording of requests entered by the user.

5. The proposed system

5.1. Architecture

The architecture of the system proposed in this section covers most of the components involved in data processing

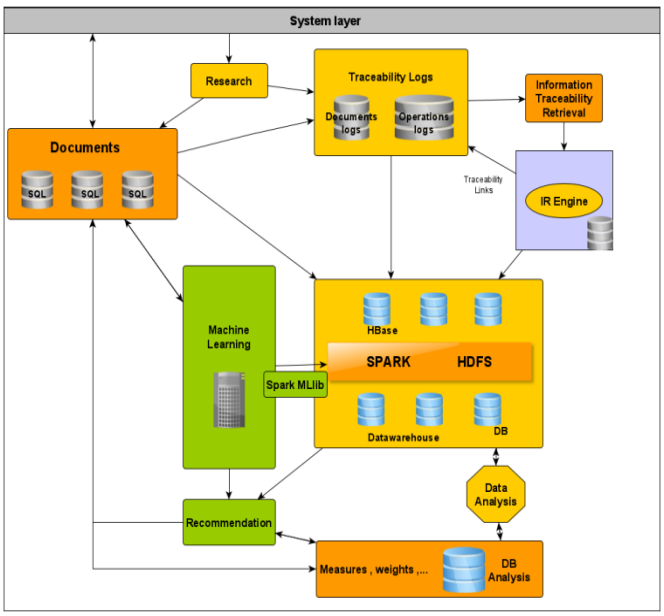

Figure 3: Architecture of the proposed system

Our system architecture consists of four main components:

- Corpus of documents: The documents coming from the corpus are heterogeneous. The user uses their interface to add new documents, even just with the action of "click and drag". Usually this is SQL data storage; (MySQL, Oracle, PostgreSQL, MongoDB, etc.). Log files will be used to manage traceability (user clicks (Action), user visits (Operation), activity, etc.).
- Data transformation: Data transformation involves the information retrieval part and converting documents into vectors, as well as data extraction so that the document can be indexed for easy searching. Indexing is a way to classify a document by adding terms to its metadata, such as tags, order numbers, or customer information. Mainly this phase is used to load the data to be processed. Logging and traceability management tools can also be considered part of data transformation, as they generate useful events from log files and present an important history to analyse.
- Data processing or integration: For data processing, this concerns the generation of usable data from a set of heterogeneous documents that can be consumed by machine learning and the big data environment; (Spark and its ecosystem (Spark / HDFS, Map-Reduce, HBase, MLlib, etc.). Spark offers in-memory scans, so this is an emergent task by reading directly from HDFS.
- Data visualization and recommendations: Visualization and recommendation represent forms of display of results and information to users or to other entities such as managers or decision-makers, in order to predict or estimate overall performance and also to generate recommendations, to using a set of algorithms, tasks defined by the user himself. (Reporting, personalized dashboards, interactive tables, statistics, research results, etc.).

5.2. Benefits and advantages

The improvement that is being made to document management systems today through this work is to offer new functionalities while keeping the classic functionalities of the system (Capture, Storage and distribution of documents):

- Traceability management: traceability management allows managers to control and track changes made to documents in the system. Traceability ensures that all actions and operations are tracked and recorded.
- Integration: Many document management systems integrate with programs such as messaging, a CRM (Customer Relationship Management) application or an ERP (Enterprise Resources Planning) database.
- Scalability: document management system should grow with the business so you don't have to change systems later.
- Security: The system must prioritize the security and protection of information and documents, as a data breach or storage problem could be disastrous for the entire organization.

- User-friendliness: The document management system should be easy to use. Users in the company should be able to easily access, manage and browse the necessary documents.
- Collaboration: The system must allow users to share and collaborate on documents.

The proposed improvement of the document management system supposed to make the work easier and more efficient, bringing many benefits: The system automates many aspects of document management. The system makes it possible to follow and see all the activities on a given document. It should apply built-in security and access controls to regularly control who can access which documents. The system ensures easy search of documents so as not to waste time and money. This makes it possible to find a document in seconds and also helps to retrieve some more relevant ones. It facilitates information sharing and collaboration, allowing documents to be accessed from multiple locations. The user can share documents, monitor workflows, grant or deny access to their documents and see what changes have been made. Also, the system offers users documents according to their profile and the history of their activities on the documents.

6. Conclusion

Using a document management system powered by machine learning, companies can modernize the way data is extracted. Such an advanced system can read information precisely to understand its intention. Based on specific models, the analysis of the document management system activated by machine learning and retrieves only the information useful to accomplish a particular job. Machine learning can also be a powerful tool for data mining and analysis in this kind of system, which will save a lot of time and effort. It is useful to examine the accuracy to 10 returned documents if one is interested in the system's ability to return relevant documents to the top of the list (which is a traditional concern of search engine users). The precision at 5, 10, 30,… returned documents nevertheless has limits: for example if a given request has only 8 relevant documents, and the processing does restore these 8 documents at the top of the list, the approach will have a precision of 10 returned documents equal to 0.8, which does not show that all the relevant documents available were found. In addition, in this example, a precision of 10 restored documents equal to 0.8 does not make it possible to determine where the two irrelevant documents are located among the ten returned. Document traceability management provides each user with proof that business processes have been respected, because if there is an error, it will easily be traced, as well as the associated time saving and improvement. Project management especially at the level of collaborative work. Several jobs are needed to improve the performance and accuracy of the document classification process. New methods and solutions are needed to obtain useful information from the growing volume of electronic documents:

- Use of semantics and ontology for document classification and information retrieval.
- For filtering and categorizing documents, the user may have folders and may want a classifier to classify each incoming document which automatically moves it to the correct folder. It is easier to find documents in sorted folders in a very large corpus.
- Reduce classifier processing and response time and improve classification accuracy, precision, and recall.
- An implementation of a meaning-based text classification procedure is required to retrieve meanings from words used in a specific context.

VSM stands out and occupies first place by its performance in all three cases. Thus, it confirms its good reputation in the literature as being one of the best classifiers. With the constant development and adaptation of machine learning, the proposed approach to document management systems will not only ensure better data analysis to meet customer needs, but also establish optimal research results. With the rapid pace at which technology and innovations invade industries with each passing day, it goes without saying that early adopters will also be the first to reap the myriad benefits that come with it.

In an electronic document management system, the majority of solutions on the market today are more concerned with the dynamic side of the system which concerns the workflow part of the document. In this work, the major objective that we tried to study is the static part by proposing for the improvement of this kind of system a multitude of techniques and approaches for this static part by attributing it a kind of fuel in order to automate it and make it as dynamic in relation to user expectations and its needs for the organization and readability of documents.

The first contribution in this direction, was the proposal of the use of traces throughout the system, thus the recovery of traces between documents in order to reduce the dimension of the treatment proposed also on documents using the techniques of information retrieval and the extraction of payload data. The voluminous mass of data has forced us to think about using a big data environment which will also help us in the analysis and distributed calculation of data.

The second contribution is also an important proposal, concerns the use of machine learning techniques in order to add to this system new types of documents (image, video, audio and text) which will be processed by proposing powerful algorithms. in this sense, and which are used to extract important data, and therefore offer users and businesses a new service in this type of system. An image or a document scanned in an electronic document management system, without it being processed manually by adding descriptions and fields to list it, it will be archived all the time. Hence the usefulness of our approach, in order to automate these methods and extract text content assembled from the metatdata of a document. This allows the IR engine and especially the indexing part which relies on the texts, the processing and the indexing of the images of the document. Because it is very easy to work on text than to work on an image. That is why we have proposed the use of content recognition algorithms which greatly assist in automatic document processing. An important point also, regarding the use of machine learning is that the preprocessing done on the images using OCR allows to extract high value textual data added during the training of an algorithm, and the The extraction of this data generates a serious increase in the volume of data, hence the aim of using a big data environment. This automated processing of multimedia documents prompted us to think of the user by offering an integrated module for recommending multimedia documents according to two criteria;

their profile settings and the history of their operations in interaction with the system.

The system proposed in this article is an improvement of the modern electronic management systems offered today. Thanks to this proposal, documents management tools also becomes easy to use, especially with an environment such as big data.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] A.W. Qurashi, V. Holmes, A.P. Johnson, "Document Processing: Methods for Semantic Text Similarity Analysis," in 2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), 1–6, 2020, doi:10.1109/INISTA49547.2020.9194665.

[2] T. Dash Roy, S. Khatun, R. Begum, A.M. Saadat Chowdhury, "Vector Space Model based Topic Retrieval from Bengali Documents," in 2018 International Conference on Innovations in Science, Engineering and Technology (ICISET), 60–63, 2018, doi:10.1109/ICISET.2018.8745587.

[3] Jian-Jun Zhou, "Study on several confidentiality protection technologies for electronic document," in Proceedings 2013 International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC), 2282–2285, 2013, doi:10.1109/MEC.2013.6885423.

[4] N. Mustafa, Y. Labiche, "The Need for Traceability in Heterogeneous Systems: A Systematic Literature Review," in 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), 305–310, 2017, doi:10.1109/COMPSAC.2017.237.

[5] C. Benkoussas, P. Bellot, A. Ollagnier, "The Impact of Linked Documents and Graph Analysis on Information Retrieval Methods for Book Recommendation," in 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 385–392, 2015, doi:10.1109/WI-IAT.2015.200.

[6] A. Boulmakoul, Z. Besri, Enterprise Organization Assessment through Structural Analysis Framework, 2013.

[7] R. Wohlrab, J.-P. Steghöfer, E. Knauss, S. Maro, A. Anjorin, "Collaborative Traceability Management: Challenges and Opportunities," in 2016 IEEE 24th International Requirements Engineering Conference (RE), 216–225, 2016, doi:10.1109/RE.2016.17.

[8] X. Jia, Z. Ma, "A Topic-based Document Retrieval Framework," in 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery, 860–864, 2012, doi:10.1109/FSKD.2012.6234372.

[9] A. Demuth, R. Kretschmer, A. Egyed, D. Maes, "Introducing Traceability and Consistency Checking for Change Impact Analysis across Engineering Tools in an Automation Solution Company: An Experience Report," in 2016 IEEE International Conference on Software Maintenance and Evolution (ICSME), 529–538, 2016, doi:10.1109/ICSME.2016.50.

[10] S. Palihawadana, C.H. Wijeweera, M.G.T.N. Sanjitha, V.K. Liyanage, I. Perera, D.A. Meedeniya, "Tool support for traceability management of software artefacts with DevOps practices," in 2017 Moratuwa Engineering Research Conference (MERCon), 129–134, 2017, doi:10.1109/MERCon.2017.7980469.

[11] S.K. Shivakumar, Digital Asset Management and Document Management, IEEE: 253–271, 2017, doi:10.1002/9781119206842.ch8.

[12] S. Irfan, B.V. Babu, "Information retrieval in big data using evolutionary computation: A survey," in 2016 International Conference on Computing, Communication and Automation (ICCCA), 208–213, 2016, doi:10.1109/CCAA.2016.7813720.

[13] A. Gandomi, M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," International Journal of Information Management, **35**(2), 137–144, 2015, doi:10.1016/j.ijinfomgt.2014.10.007.

[14] Big Data, Fast Data and Data Lake Concepts, Procedia Computer Science, **88**, 300–305, 2016, doi:10.1016/j.procs.2016.07.439.

[15] Introduction au Machine Learning, Machine Learnia, 2019.

[16] S. Gopal, S. Raghav, "Automatic document retrieval using SVM machine learning," in 2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon), 896–901, 2017, doi:10.1109/SmartTechCon.2017.8358501.

[17] F.O. Isinkaye, Y.O. Folajimi, B.A. Ojokoh, "Recommendation systems: Principles, methods and evaluation," Egyptian Informatics Journal, **16**(3), 261–273, 2015, doi:10.1016/j.eij.2015.06.005.

[18] T. Badriyah, S. Azvy, W. Yuwono, I. Syarif, "Recommendation system for property search using content based filtering method," in 2018 International Conference on Information and Communications Technology (ICOIACT), 25–29, 2018, doi:10.1109/ICOIACT.2018.8350801.

[19] Y. Afoudi, M. Lazaar, M. Al Achhab, "Impact of Feature selection on content-based recommendation system," in 2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), 1–6, 2019, doi:10.1109/WITS.2019.8723706.

[20] J.S. Mary, S. Usha, "Web based document management systems in life science organization," in 2015 Online International Conference on Green Engineering and Technologies (IC-GET), 1–3, 2015, doi:10.1109/GET.2015.7453826.

[21] T.V. Khronusova, S.V. Kruchinin, E.V. Bagrova, "Implementation of Electronic Document Management in Russian Education. Quality Assessment," in 2019 International Conference "Quality Management, Transport and Information Security, Information Technologies" (IT QM IS), 608–610, 2019, doi:10.1109/ITQMIS.2019.8928356.

[22] S.V. Kruchinin, E.V. Bagrova, "Systems of Electronic Document Management in Russian Education. Pros and Cons," in 2019 International Conference "Quality Management, Transport and Information Security, Information Technologies" (IT QM IS), 628–630, 2019, doi:10.1109/ITQMIS.2019.8928315.

[23] K. Souali, O. Rahmaoui, M. Ouzzif, "An overview of traceability: Definitions and techniques," in 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt), 789–793, 2016, doi:10.1109/CIST.2016.7804995.

[24] S.F. Königs, G. Beier, A. Figge, R. Stark, "Traceability in Systems Engineering – Review of industrial practices, state-of-the-art technologies and new research solutions," Advanced Engineering Informatics, **26**(4), 924–940, 2012, doi:10.1016/j.aei.2012.08.002.

[25] D. Asioli, A. Boecker, M. Canavari, "Perceived Traceability Costs and Benefits in the Italian Fisheries Supply Chain," International Journal on Food System Dynamics, **2**(4), 357–375, 2011, doi:10.18461/ijfsd.v2i4.242.

[26] H. Cui, Z. Chen, Y. Xi, H. Chen, J. Hao, "IoT Data Management and Lineage Traceability: A Blockchain-based Solution," in 2019 IEEE/CIC International Conference on Communications Workshops in China (ICCC Workshops), 239–244, 2019, doi:10.1109/ICCChinaW.2019.8849969.

[27] X. Chen, J. Hosking, J. Grundy, "A combination approach for enhancing automated traceability: (NIER track)," in 2011 33rd International Conference on Software Engineering (ICSE), 912–915, 2011, doi:10.1145/1985793.1985943.

[28] S.M. Bakhshavesh, A. Mohebil, A. Ahmadi, A. Badarnchi, "A New Subject-Based Document Retrieval from Digital Libraries Using Vector Space Model," in 2018 Federated Conference on Computer Science and Information Systems (FedCSIS), 161–164, 2018.

[29] E. Wahyudi, S. Sfenrianto, M.J. Hakim, R. Subandi, O.R. Sulaeman, R. Setiyawan, "Information Retrieval System for Searching JSON Files with Vector Space Model Method," in 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT), 260–265, 2019, doi:10.1109/ICAIIT.2019.8834457.

[30] D.V. Rodriguez, D.L. Carver, "Multi-Objective Information Retrieval-Based NSGA-II Optimization for Requirements Traceability Recovery," in 2020 IEEE International Conference on Electro Information Technology (EIT), 271–280, 2020, doi:10.1109/EIT48999.2020.9208233.

[31] S. Pandey, I. Mathur, N. Joshi, "Information Retrieval Ranking Using Machine Learning Techniques," in 2019 Amity International Conference on Artificial Intelligence (AICAI), 86–92, 2019, doi:10.1109/AICAI.2019.8701391.

[32] Q. Shuai, R. Wang, L. Jin, L. Pang, "Research on Gender Recognition of Names Based on Machine Learning Algorithm," in 2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), 335–338, 2018, doi:10.1109/IHMSC.2018.10182.

[33] P. Bhavsar, I. Safro, N. Bouaynaya, R. Polikar, D. Dera, Chapter 12 - Machine Learning in Transportation Data Analytics, Elsevier: 283–307, 2017, doi:10.1016/B978-0-12-809715-1.00012-2.

[34] S. Solanki, S. Verma, K. Chahar, "A Comparative Study of Information Retrieval Using Machine Learning," in: Sharma, H., Govindan, K., Poonia, R. C., Kumar, S., and El-Medany, W. M., eds., in Advances in Computing and Intelligent Systems, Springer, Singapore: 35–42, 2020, doi:10.1007/978-981-15-0222-4_3.

[35] C. Goller, J. Löning, T. Will, W. Wolff, Automatic Document Classification - A thorough Evaluation of various Methods, Undefined, 2000.

[36] L. Getoor, Link-based Classification, Springer, London: 189–207, 2005, doi:10.1007/1-84628-284-5_7.

[37] C. Mills, J. Escobar-Avila, S. Haiduc, "Automatic Traceability Maintenance via Machine Learning Classification," in 2018 IEEE International Conference on Software Maintenance and Evolution (ICSME), 369–380, 2018, doi:10.1109/ICSME.2018.00045.
[38] R. Oliveto, M. Gethers, D. Poshyvanyk, A. De Lucia, "On the Equivalence of Information Retrieval Methods for Automated Traceability Link Recovery," in 2010 IEEE 18th International Conference on Program Comprehension, 68–71, 2010, doi:10.1109/ICPC.2010.20.
[39] J.H. Hayes, A. Dekhtyar, J. Osborne, "Improving requirements tracing via information retrieval," in Proceedings. 11th IEEE International Requirements Engineering Conference, 2003., 138–147, 2003, doi:10.1109/ICRE.2003.1232745.
[40] K.K. Agbele, E.F. Ayetiran, K.D. Aruleba, D.O. Ekong, "Algorithm for Information Retrieval optimization," in 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 1–8, 2016, doi:10.1109/IEMCON.2016.7746242.
[41] (PDF) Algorithm for calculating relevance of documents in information retrieval systems | IRJET Journal - Academia.edu, Nov. 2020.
[42] Information retrieval models for recovering traceability links between code and documentation - IEEE Conference Publication, Oct. 2020.
[43] S. Nagano, Y. Ichikawa, T. Kobayashi, "Recovering Traceability Links between Code and Documentation for Enterprise Project Artifacts," in 2012 IEEE 36th Annual Computer Software and Applications Conference, 11–18, 2012, doi:10.1109/COMPSAC.2012.10.
[44] C. Mills, S. Haiduc, "A Machine Learning Approach for Determining the Validity of Traceability Links," in 2017 IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C), 121–123, 2017, doi:10.1109/ICSE-C.2017.86.
[45] G. Spanoudakis, A.S. d'Avila Garcez, A. Zisman, "Revising Rules to Capture Requirements Traceability Relations: A Machine Learning Approach," in Proceedings of the Fifteenth International Conference on Software Engineering & Knowledge Engineering (SEKE'2003), Hotel Sofitel, San Francisco Bay, CA, USA, July 1-3, 2003, 570–577, 2003.
[46] I.D. Schizas, "Online Data Dimensionality Reduction and Reconstruction Using Graph Filtering," IEEE Transactions on Signal Processing, **68**, 3871–3886, 2020, doi:10.1109/TSP.2020.3003423.
[47] J. Laaksonen, E. Oja, "Classification with learning k-nearest neighbors," in Proceedings of International Conference on Neural Networks (ICNN'96), 1480–1483 vol.3, 1996, doi:10.1109/ICNN.1996.549118.
[48] K. Taunk, S. De, S. Verma, A. Swetapadma, "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification," in 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 1255–1260, 2019, doi:10.1109/ICCS45141.2019.9065747.
[49] KNIME | Open for Innovation, Oct. 2020.
[50] O. Harrison, Machine Learning Basics with the K-Nearest Neighbors Algorithm, Medium, 2019.