I.J. Intelligent Systems and Applications, **2022, 2, 1-13**
Published Online April 2022 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2022.02.01

Towards an Efficient Big Data Indexing Approach under an Uncertain Environment

Asma Omri
MARS Research Laboratory LR17ES05, University of Sousse, Tunisia
E-mail: omri.asmaaa@gmail.com

Mohamed Nazih Omri
MARS Research Laboratory LR17ES05, University of Sousse, Tunisia
E-mail: mohamednazih.omri@eniso.u-sousse.tn

Received: 07 October 2021; Revised: 18 November 2021; Accepted: 02 January 2022; Published: 08 April 2022

Abstract: It is generally accepted that data production has experienced spectacular growth for several years due to the proliferation of new technologies such as new mobile devices, smart meters, social networks, cloud computing and sensors. In fact, this data explosion should continue and even accelerate. To find all of the documents responding to a request, any information search system develops a methodology to confirm whether or not the terms of each document correspond to those of the user's request. Most systems are based on the assumption that the terms extracted from the documents have been certain and precise. However, there are data in which this assumption is difficult to apply. The main objective of the work carried out within the framework of this article is to propose a new model of data service indexing in an uncertain environment, meaning that the data they contain can be untrustworthy, or they can be contradictory to another data source, due to failure in collection or integration mechanisms. The solution we have proposed is characterized by its Intelligent side ensured by an efficient fuzzy module capable of reasoning in an environment of uncertain and imprecise data. Concretely, our proposed approach is articulated around two main phases: (i) a first phase ensures the processing of uncertain data in a textual document and, (ii) the second phase makes it possible to determine a new method of uncertain syntactic indexing. We carried out a series of experiments, on different bases of standard tests, in order to evaluate our solution while comparing it to the approaches studied in the literature. We used different standard performance measures, namely precision, recall and $\boldsymbol{F_{measure}}$. The results found showed that our solution is more efficient and more efficient than the main approaches proposed in the literature. The results show that the proposed approach realizes an efficient Big Data indexing solution in an Uncertain Environment that increases the Precision, the Recall and the $\boldsymbol{F_{measure}}$ measurements. Experimental results present that the proposed uncertain model obtained the best precision accuracy 0.395 with KDD database and the best recall accuracy 0.254 with the same database.

Index Terms: Indexing, Probabilistic, Big Data, Syntactic, Uncertain.

1. Introduction

Information retrieval (IR) has long attracted the attention of the scientific community and is generally defined by the identification of documents that best meet a user's need for information. These documents should be found among a large collection of unstructured documents [1, 2] With the expansion of the internet, the implementation of solutions capable of exploiting web content and improving search performance has become essential. Techniques have been proposed and applications have been carried out, their objective is to provide users with answers that are relevant to the needs they express. For this reason, one of the axes which should be the most interesting is that of indexing [3, 4, 5].

Indexing plays an important role in the field of information retrieval. There are several types of indexing, but in general, the principle is to convert the data sources into electronic data. Indexing methods have been widely studied and are a very important option for Big Data [6] Most of the research proposed in the literature considers that indexing is already defined on a single data problem. The data that exists today presents a set of challenges and risks that researchers want to address [6]. One of these challenges is the uncertainty of data sources and data.

In this article we therefore question the veracity of this basic equality in certain contexts: uncertain data. The presence of uncertainty in the data questions this equality. In an uncertain context, terms have been poorly recognized or identified implying poor performance of information retrieval systems. It is therefore necessary to rethink an information retrieval system in order to adapt it to this type of data. In order to overcome this problem, we propose a new information retrieval method which is capable of processing uncertain data based on a probabilistic approach. Our main

challenge is to design a highly efficient indexing mechanism that can withstand today's data characteristics and above all data uncertainty.

1.1. Challenges

In this section, we will present the model we proposed for indexing data, taking into account the treatment of uncertainty or the veracity of the data. In this section, we focus on representing and dealing with uncertainties about the value a variable can take. This uncertainty can come from the intrinsic variability of the phenomena influencing the value of this variable, the inaccuracy or unreliability of the variables. We focus on the indexing phase in the presence of uncertain data. To reach our goal, we must answer the following questions:

- **How to calculate the uncertainty of the terms:** the establishment of a new model of representation of the terms remains indispensable. This model must be able to determine the degree of certainty of each term extracted from a document in the information search domain. It will be articulated around a module based on the theory of probabilities where each term will be associated with a value P which expresses its degree of accuracy within a specific document.
- **Develop a module for calculating the degree of uncertainty in the syntactic indexing part:** a new syntactic indexing method, in the form of a module, must be proposed to cope with the uncertain data. This method must be able to return all the terms extracted from a document taking into account their probability values. It is not a question of returning these values of uncertainty but rather of proposing a method which makes it possible to build the indexed document in an uncertain environment.

In what follows, we attempt to detail the principle of each stage in our conceptual model.

1.2. Motivation of the proposed approach

In order to build our proposal, we rely on a scenario of an application in the field of information retrieval. To search for all documents responding to a request, the information retrieval system relies on a formal or operational methodology to check the correspondence between the terms of each document or the terms of the user's request. Most systems object that the component terms of documents have been fully recognized or identified. The step that deals with the extraction of these terms is the indexing step, where the fundamentals of the systems require that the terms extracted from the documents are certain terms. After the development of technologies and social networks, the data become uncertain for several reasons. The example shown in Figure 1, [7] in which the information on audio documents (conversations for example) are extracted by speech recognition tools. These tools provide for each audio document a sequential list of words related to the trust of their recognition. To treat these documents is therefore to treat documents whose words are uncertain and associated uncertainties related to the extraction process. Beyond the extraction of words, these systems address the problem of the ambiguity of words: to know if 'door', for example, is a certain or uncertain word in a document makes it possible to ensure a better accuracy in the face of a problem. Query specifying only one of the two possibilities. From this example, we find that there are indeed contexts in which the hypothesis that a document of the corpus is a sequence of certain words. In these contexts, the document must be considered as a sequence of words associated with extraction hypotheses. As a result, the information retrieval system that relies on such documents must incorporate this new dimension that allows us to determine whether extracted words are certain or uncertain words. The purpose of this article is to show a first approach where we first propose an approach for syntactic indexing in the presence of uncertain data, a semantic indexing approach in an uncertain environment and a hybrid approach for data indexing, especially taking into account data uncertainty.

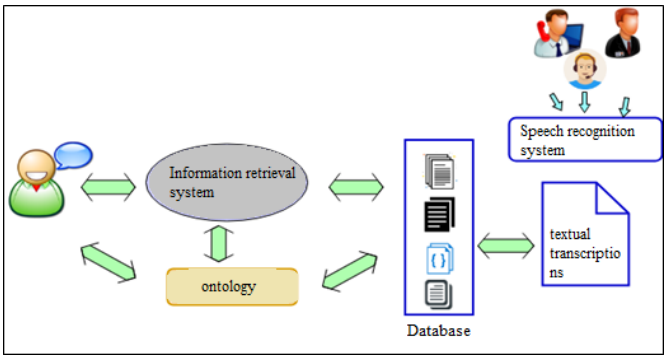

Fig.1. Example of motivation

Figure 1 presents a tool for exchange between different users of such an application can be made using two types of documents namely written and oral documents. Automatic conversation transcriptions are uncertain semantic data.

An information retrieval system capable of managing data uncertain is necessary in this type of application. In this context, we will focus on the indexing of information retrieval system that will be adapted to uncertain data.

1.3. Contributions

In this article, we propose a probabilistic approach for the representation of the terms of a document in the field of information retrieval. We will focus on the indexing phase of data in an uncertain environment. Below, we summarize our main contributions:

- **Modeling approach of the terms contained in a textual document:** A method of representing uncertain terms extracted from a document based either on a possibilistic approach or a probabilistic approach. For the sake of simplicity, our method is based on the theory of probability or for each term t, belonging to a document D, is associated a degree of probability. This degree expresses how this term indexes the document D in a request.
- **Uncertain syntactic indexing approach:** To process uncertain data at the information retrieval domain level, a new syntactic indexing method has been proposed. This method takes into account the degrees of probability associated with each term. This value must be either involve at the weighting phase of the terms, or use in a new equation.

1.4. Structure of the paper

The rest of this paper is structured as follows. The next section 2 describes the different basic notions of our proposition. In section 3, we describe the principle of calculation of uncertainty. Section 4 presents the motivation of the proposed approach. Section 5 presents the study and analysis of algorithm complexity indexing. Section 6 describes the results of the experimental study and details the evaluation of our approach. Section 7 gives a general discussion on the main results of our proposed approach. In section 8, we summarize the main limits of our work. Section 9 concludes this paper and gives some perspectives for the different study's contributions.

2. Related Works

By reviewing the literature, we found that many uncertainty approaches have been proposed. Each approach is characterized by its own criteria and has been classified into two categories: (i) a first category which deals with Uncertainty in the web and (ii) a second category which deals with Uncertainty in indexing.

2.1. Uncertainty in web

The first model [1] that we studied, in the literature, its authors proposed a new approach, which dynamically determines the effect of the evaluation factors based on an interactive Big Data system by integrating an inference module fuzzy (FIS). The fuzzy multi-factor assessment (MFE) approach was applied to be able to compare and evaluate interactive big data systems with each other. The second model [8] studied, its authors proposed a parallel sampling method based on the notion of Hyper Surface dedicated to big data with uncertainty distribution. For example, PSHS adopts a universal concept of minimum coherent subset (MCS) of the Hyper Surface Classification (HSC). These authors were able to manage the uncertainties of sampling from the big data. PSHS was then implemented on the basis of the MapReduce framework, which represents a parallel programming technique widely used today given its performance in many areas. In this research [9, 10], they proposed an ontology-based method to solve an uncertainty problem for large data sources. In model [11], the authors focus on the fourth V, namely veracity, to demonstrate the essential impact of uncertainty modeling on the improvement of learning performance. Low veracity is a change in uncertainty and missing values on a large scale. With the exception of the altered uncertainty of the data itself, the uncertainty in the modeling and processing of the data also changes in very remarkable ways.

2.2. Uncertainty in indexation

A new strategy approach was presented in the work of [12] and is based on the DRASTIC framework. This approach integrates three main modules: (i) a module for the reduction of subjectivities; (ii) a second module to transform the indexing of vulnerabilities into indexing of risks; and a third module to understand the inherent uncertainties. The notion of indexing refers to relative values for spatial comparisons within an aquifer using a set of integrated data layers. Also, the risk indexing reveals the fact that the study area suffers from some inappropriate management practices both in terms of agricultural activities and in terms of groundwater catchment. In the work of [13], the authors propose a new more or less efficient method of intermediate indexing, called MI, which makes it possible to filter a large quantity of irrelevant data in the uncertain data stream. This filtering is done by specifically sorting the data, in order to improve the efficiency of updating the k-dominant horizon. Another approach has been proposed by [14] in which the authors try to solve the problem of classifying uncertain objects whose locations are described by probability density functions (pdfs). These authors have thus proposed pruning techniques based on Voronoi diagrams to reduce the number of expected distance calculations. They then introduced an R-tree index to organize uncertain objects which will help reduce

pruning costs. The authors of the approach presented in [15], proposed a new structure to capture the probability density function of an uncertain object, called MRST. Considering the gradient of the probability density function, this new MRST structure could provide an uncertain object with high pruning power and thus consume less space. At the same time, they proposed a new index named R-MRST to efficiently support range queries on multidimensional uncertain data. This index has a strong power of size and a low cost in terms of space and dynamic update. In the work presented by [16], the authors demonstrate a new way of constructing an efficient index structure to handle large queries, of uncertain range and which presents a certain similarity. To do this, they have developed query processing techniques that use efficient index search methods to calculate end results. The approach proposed by [17], allows to examine the traditional systems for the explicit management of the uncertainty present in the data. They thus proposed two index structures to efficiently search for uncertain categorical data, one based on the R tree and the other based on an inverted index structure. A detailed description of the probabilistic equality queries they support has also been presented. The work presented by [18], describes the uncertain trajectory model of moving objects in road network databases and extends the structure of the MON-tree index according to this network. The uncertain geometry between two sample points is split into chunks and then stored in the list of uncertain areas based on the time sorting. Another approach has been proposed by [19]. This approach supports the fact that the final round functions as an epistemic downgrade by (a) making non-problematic non-confirmation possible afterwards, (b) gesturing towards an unverbalized alternative, and (c) being oriented format. This work contributes to several major areas of conversation analysis research, mainly interactive grammar and epistemics.

In some previous studies [20], a probability-based method for modeling and indexing uncertain spatial data is proposed. In this paper, they represented uncertain objects with PDFs (probability density function) and defined a similarity measure for these uncertain objects. To index the objects, they designed an Optimized Gaussian Mixture Hierarchy (OGMH) to support both certain/uncertain queries and certain/uncertain data. They used for the particular case (certain query) an uncertain R tree with two query filtering schemes, UR1 and UR2. Finally, they applied the similarity measure, the uncertainty model and the indexing structures on a set of data.

Some other research [21] proposed the SQUiD framework, which combines an index-based technique with a prediction method to reduce the computation time to compute horizons on uncertain high-dimensional data. Through experiments, the results clearly demonstrate the effectiveness of the proposed algorithm compared to prior findings.

Networks in many real-world applications have inherent uncertainty in their structure. This article [22] highlights the problem of indexing and querying (k,γ)-lattices on uncertain graphs. To keep the complete information of (k,γ)-lattice, they proposed a compact data structure of the CPT index. They proposed two index building algorithms, CPT-Basic and CPT-Fast. To efficiently build the CPT index using top-down graph partitions, they developed a bottom-up CPT index building scheme and an improved algorithm. Based on CPT-Fast, they developed a scheme to find a compromise between index building and online retrieval processes. Extensive experiments on large datasets have shown the superiority of our indexing methods over other state-of-the-art approaches.

The treatment of uncertainty in the field of indexing and the extraction of relevant information from the web, under different data modeling techniques, has also been mainly addressed in the following works: Omri's work [23], for example, proposes a feedback model on relevance for goal extraction from fuzzy semantic networks. On the other hand, the authors of [24] propose another new model of information retrieval based on possibilistic Bayesian networks. The third model, presented in this context, is that of [25]. In their work, the authors describe their solution which consists of a fuzzy new approach to extracting relevant information from web resources. The last model that we have studied, in this perspective, is the one quotes [26] where the author proposes a feedback model of possibilistic relevance and semantic networks for the extraction of goals.

3. Probability Theory: Background

Artificial intelligence communities and databases have been rarely treated especially by management of uncertainty. It has also received a very little attention in the field of service-oriented architectures [27, 28]. Data uncertainty can be generated from local sources to which different types of imperfection are involved (e.g., imprecision, uncertainty, incompleteness, inconsistency, etc.). Literature has also stated that data available for an information system is often imperfect [29]. Information is perfect if it is precise and certain. The reasons behind imperfection are notably uncertainty, inaccuracy and inconsistency. These last concepts represent the major aspects of the imperfect data [30, 31] which will be defined immediately after. Generally, vagueness and inconsistency are linked to the content of the information. These concepts are associated to the content of information while uncertainty is linked to the validity of the information. Uncertainty results from lack of information about such a situation, inconsistency between information, incompleteness of information and variability.

The theory of probability was used in the analytical essays of the game of Pierre de Fermat and Blaise Pascal in the 17th century and that of Gerolamo Cardano in the 15th century. In 1657, Christiaan Huygens published a book on this subject and in the 19th century, Pierre Laplace, completed the notion of classical interpretation [32]. Probability theory therefore provides a solid representational language and a computational environment for rational degrees of belief. This allows different communities the free choice of having different beliefs on the same given hypothesis. This

is considered to be the result of a random event which cannot be determined until it occurs, but it can be one of many possible results [33]. From here we can say that what has been put forward provides a compelling framework for representing uncertain and imperfect knowledge that comes from various sources. In the literature, there are several approaches that use probabilities in the domain of the Semantic Web [34, 35, 36]. The key objects of probability theory are random variables, stochastic processes and events: three essential concepts for the modeling of our solution.

4. Problem formalization

Most approaches are based on the assumption that terms extracted from documents have been fully recognized or identified [37, 38, 39, 40]. In some cases, the terms extracted are uncertain, that is, their identification is uncertain [1]. As a result, the equality relation between the term of the document and the term of the query is no longer true. This raises the question of the behavior of an information retrieval system under such conditions. We were interested, within the framework of our work, in the representation and the analysis of the uncertain data that a variable can take. This uncertainty can come either from the intrinsic variability of the phenomena influencing the value of this variable, or from the imprecision or unreliability of the information available. Our goal is to focus on the indexing phase in the presence of uncertain data. To reach this objective, and thus propose an effective solution, we will proceed by, as a first step, to make a comparative study between the main methods of uncertainty treatment proposed in the literature [4]. We can cite: (i) probability theory, (ii) uncertainty theory, (iii) fuzzy logic, (iv) possibility theory, (v) P-boxes, and (vi) Dempster-Shafer theory. In our approach we will be interested in the uncertainty at the language level. This principle is based on the context of "language model".

Another way to model documents in probabilistic form is in the use of language models. These language models provide a representation of a language. This representation can be used within information retrieval models. The principle of language models is to determine the certainty of each word or phrase based on probability theory. This language model assigns a probability to each expression appearing in a document. Language models also play a central role in statistical machine translation [1]. A language model is constructed using "a probability function P that assigns a probability P(S) to a word or sequence of words S in a language" [4, 41] (see Figure 2). We call language a body of documents. This function makes it possible to estimate the probability of any sequence of words in the modeled language or, more generally, to estimate the probability of generating this sequence of words from the language model.

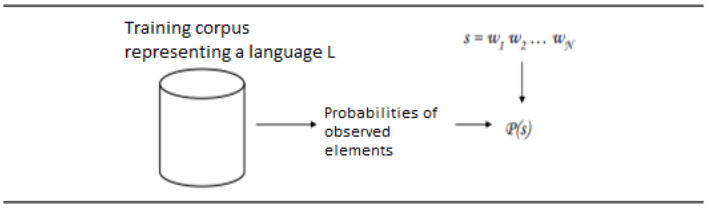

Fig.2. Principle of operation of language models

4.1. Notations

In this section, we will present the main notations that will be used in this article:

- D: Document
- C: Corpus
- N = number of documents in the corpus
- tf (t, d) = term frequency = number of occurrences of the term t in the document d
- idf (t) = inverse document frequency = the number of documents indexed by the term t in a collection
- Cf: the concept frequency
- P: degree of probability

4.2. Calculation of uncertainty

To calculate the probability of belonging to the language L of a sentence presented by the following expression: Ph = $m_1\ m_2\ \dots m_n$, we have to use the following equation which allows us to estimate the probability of the sentence to be in the ML model of the language [43]:

$$\boldsymbol{P(Ph) = \prod_{i=1}^{n} P_{ML}(M_i | M_1, \dots, M_{i-1})} \quad (1)$$

We use a corpus of documents that allows us to represent the language to model [42, 43, 44] to estimate the probability value. The probability of a term m in a document corpus C is based on the maximum likelihood estimate of the word m and it is given by the next equation:

$$\boldsymbol{P_{ML}(\mathbf{M}) = \frac{|\mathbf{M}|}{\sum \mathbf{M} \in \mathbf{C}} = \frac{\mathbf{NTC}}{\mathbf{NN-GC}}} \quad (2)$$

Where:

NOTC: represents the Number of occurrences of term M in the corpus C.

NNGC: represents the Number of N-Gram of C.

For the degree of the uncertainty of words extracted from a document, is based on the principle of information search language model.

Certainty calculation of a term: Let a certainty value c associated with each term a ∈ V_c. We represent this value by the following Fcert certainty function Fcert: $V_c \rightarrow R^+$

$$a \rightarrow P$$

To better understand the principle, take the following sentence as an example: Ph = "Isabelle is in the spotlight. This honor ...". We assume that we have a data extraction process such as an automatic speech recognition system that outputs a sentence written in a document. The general model principle that we will propose is presented by the following figure:

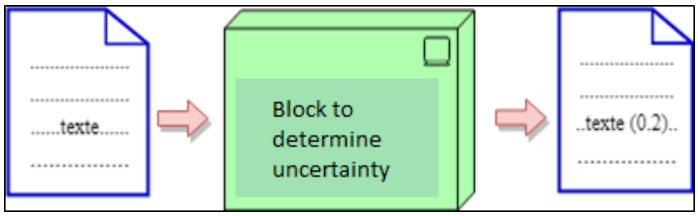

Fig.3. Uncertainty calculation principle

After taking the Ph input, this model gives us as output:

Pexit = « Isabelle is a hooker. This honor ... » With the following indications:

- Cert (Isabelle) = 0.7
- Cert(is) = 0.2
- Cert(hooker) = 0.5
- Cert(This) = 0.18
- Cert(honor) = 0.8

O.7, 0.2, 0.5, 0.18, 0.8 correspond respectively to the uncertainty values associated with the terms *isabelle, is, hooker, this, honor*. In what follows in this paper, we will explain the principle of each indexing approach in an uncertain environment.

5. Proposed Indexing Big Data Approach

5.1. Architecture of the proposed model

This part aims to add an uncertainty calculation stage in the syntactic indexing phase. The principle of this method is presented by the following figure:

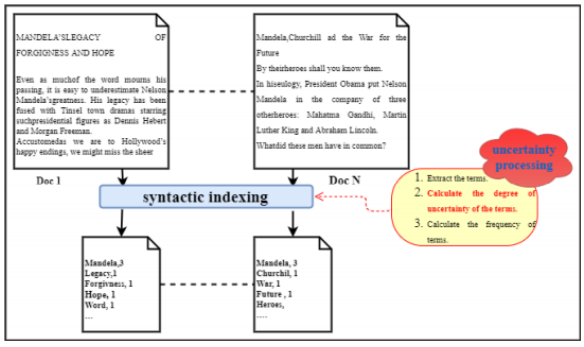

Fig.4. Principle of uncertain syntactic indexing

This approach is based essentially on the standard indexing principle which takes as input a set of textual documents and gives us a set of indexed documents. After extracting the terms of the documents, we try to calculate the uncertainty degree of these terms by using the uncertainty calculation method ($Incert_{ij}$: represents the degree of uncertainty of term i in a document j) explained in the previous part. This technique measures a degree of uncertainty in all terms i in a document j. Finally, we try to calculate the frequency of each term. Then we must either concatenate the uncertainty value of each term and the weight of each term is to find a relationship between weight and degree of uncertainty.

5.2. Indexing textual documents algorithms

Algorithms 1 and 2 describe the principle of indexing textual documents in the presence of uncertain data based on an approach. It makes it possible to effectively return all the words presented in a document by calculating their degree of probability which determines their certainties. The goal of this probabilistic syntactic indexing is to extract the terms from documents and store them with their number of occurrences and degrees of certainty in a physical index.

Algorithm 1: Probabilistic document

```
Data: D : document.  C : Corpus.
Result: D_Index_Syn_Pro : Probabilistic documents.
begin
    for Each document D in Corpus C do
        2. /*for all documents D in the corpus, D is the current document*/ Create(DPro) /* Creation of the probabilistic document*/
        for every line L in D do
            */For all lines of the current document DPro, Current Line: L*/
            for Every word w in L do
                /* for all the words of a line L*/
                if fin_mot(w) == false then
                    /* if the word is not equal to the end_mot*/
                    if DPro.Contain(w) == false then
                        /*Le document probabiliste ne contient pas le mot w*/
                        DPro.Add(W,Cert(t))
                        /* add the word to the probabilistic document with a word uncertainty value p = Cert(t) */
                    end
                    else
                        /*The probabilistic document contains the word w, then Update the value of P*/
                        DPro.UpdateP(P * w.P)
                    end
                end
            end
        end
    end
end
```

Algorithm 2: Probabilistic syntactic indexing

```
Data: DPro : Probabilistic document.  C : Corpus.
Result: D_Index_Syn_Pro : Probabilistic syntactic indexing documents.
begin
    for Every DPro document in Corpus C do
        /*for all DPro documents in the corpus, DPro is the current document*/ Create(D_Ind_Synt_Pro) /*Create (Probabilistic syntactic
          indexation Ind_Synt_Pro of the current document D*/
        for Each line L in DPro do
            */For all lines of the current document DPro, Current Line: L*/
            for every word w in L do
                /* for all the words of a line L*/
                if end_mot(w) == false then
                    /* if the word does not equal theend_mot*/
                    if D_index_syn.Pro.Contain(w) == false then
                        /*The probabilistic syntactic indexing document does not contain the word w*/
                        D.Index_Syn_Pro.Add(W, 1, 1)
                        /* add the word to the probabilistic syntactic indexing document with a frequency of word tf = 1 an uncertainty Incert
                          =1*/
                    end
                    else
                        /*Index_Syn contain the word: increment tf*/;
                        /* The probabilistic syntactic indexing document contains the word w*/
                        /* increment tf*/
                        D. Index_Syn_Pro.IncreTF(w.Tf)
                        /*Extract the probability value of each term Incert*/
                        /* Change the uncertainty value */
                        D.Index_Syn_Pro.UpdateP(P * Incert)
                    end
                end
            end
        end
    end
end
```

These algorithms are divided into 3 parts: (i) The first algorithm consists in building the set of probabilistic documents. This phase takes as input a set of documents of corpus, one goes through all the documents of corpus. For each document, one must go through all the lines of each document. For each line, one must go through all the terms of each line. For each term one must determine the degree of probability. The result is a set of documents where each term is characterized by its degree of certainty. (ii) The second phase is presented in the second algorithm (lines 1-14) of index construction, one must extract the terms meaning from the document jump the term that indicates the end of term. (iii) The third phase in the second algorithm (lines 14-27) consists of calculating the occurrence of each term and the probability of each term. For each term signifier, the frequency of occurrence must be calculated, and the probability must be determined. Whenever we find a term, we increment the frequency of this term, we make the product of probabilities. The result of this phase is an index-Syn_Pro file for each document.

For each document, one obtains all the extracted terms of documents belonging to a corpus with their values of frequency and their degrees of probability:

$$\boldsymbol{Document_{ID}} \Rightarrow \{(\boldsymbol{term_1}, \boldsymbol{tf_1}, \boldsymbol{P_1},), \dots, (\boldsymbol{term_n}, \boldsymbol{tf_n}, \boldsymbol{P_n})\} \quad (3)$$

With:

- n: number of significant words of document D.
- P: probability of each term.
- Tf: frequency of occurrence of each term.

5.3. Study and analysis of algorithm complexity indexing

In this part, we study the complexity of the probabilistic syntactic indexing algorithm by analyzing the complexity of its different phases.

- In the first part, each textual document D is transformed into a probabilistic textual document. Suppose N is the number of documents to index in C (ie a document collection), LMax is the maximum number of Lines per document and MMax is the maximum number of possible words per line belonging to a document D. The complexity of the first phase is given by the quantity O (N * LMax * MMax).
- In the second part, the Dpro probabilistic documents are already recovered. This phase is interested in the construction of indexes, we must extract the terms meaning of the document except the term which indicates the end of term. A browse of all these documents is necessary for indexed.
- The third phase consists of calculating the occurrence and probability of each term. For each term signifier, the frequency of occurrence must be calculated, and the probability must be determined. Whenever a term is found, we increment its frequency this, we then calculate the product of the probabilities. We assume that we have an indexed document DI., Each document contains LDIMax lines and line contains MDIMax words, so the complexity of this phase is given by, O (LDIMax * MDIMax).

In summary, the different phases of our algorithm have a polynomial complexity.

6. Experimental Study and Results Analysis

In this section, we will perform a series of experiments in order to validate our proposed approach and confirm its efficiency in processing data in an uncertain environment. We also want, through this series of experiments, to show that in terms of execution time our solution is better than that of the main standard methods studied in the literature. To do this, we implemented the models: Standard Syntax Indexing Method and our Uncertain Syntax Indexing Model (ISI). We have implemented our algorithms in the Java programming language and under the Windows 7 environment. To evaluate the performance in runtime, experimental tests were carried out on a Pentium (R) Dual Core machine with a clock frequency of 2.2 GHz and 3 GB of main memory.

6.1. Description of the test corpora collection

The main objective that we have set for ourselves is to carry out a comparative study between the different models (taking the execution time as an indicator) while varying the size of the data (size of the corpus) of the test. To achieve this goal, we have prepared a set of corpora each containing a number of attribute files ranging from 500 to 20,000.

We then generated a network corresponding to each corpus. In practice, this network is stored in an XML file because the tree structure provided by the XML language allows easy data retrieval. Note that the XML file is created by Lattice Navigator 3, a tool for generating and visualizing concept networks.

6.2. *Performance metrics used for the evaluation*

In the following, we present the precision, recall and $\boldsymbol{F_{measure}}$ which we will use to evaluate the performance of our proposed algorithms. These measures are defined based on the next confusion matrix.

Table 1. Diagram of confusion matrix

	Predicted positive sample	**Predicted negative sample**
True positive sample	TP	FN
True negative sample	FP	TN

The evaluation measures used, are precision (also called positive predictive value) which is the fraction of relevant instances among the retrieved instances, recall (also known as sensitivity) which is the fraction of relevant instances that have been retrieved over the total amount of relevant instances, $F_{MEASURES}$ which is a combination between precision and recall. These evaluation measures are calculated respectively using Diagram of confusion matrix presented in Table 1 and equations 4, 5, and 6.

$$\boldsymbol{Precision} = \frac{\mathbf{TP}}{\mathbf{TP+FP}} \tag{4}$$

$$\boldsymbol{Recall} = \frac{\mathbf{TP}}{\mathbf{TP+FN}} \tag{5}$$

$$\boldsymbol{F_{measure}} = \mathbf{2} * \frac{\boldsymbol{Precision*Recall}}{\boldsymbol{Precision+Recall}} \tag{6}$$

6.3. *Experimentation and results*

In practice, we calculate for each model the response time to a given request. We estimate the execution times needed for indexing with and without calculating the degrees of probability. Figure 5 exposes the evolution of the execution time for each model as a function of the number of objects in the network.

This study proves, experimentally, the importance of the theoretical results presented in the curves. Compared to the standard model, our approach ISI has a low time complexity which negatively affects the execution time but positively affects the quality of the response. Figure 5 shows that Our approach ISI curve is slightly higher than the other curve, which is considered as a good indicator of speed. If we classify the models according to their speed, we obtain the following order: Standard indexing model, Our approach ISI. This result seems logical to us, because the part which we proposed to calculate the correctness of the terms takes a little time on the other hand, the model which we proposed greatly improves the quality of the answer.

Fig.5. Performance results in terms of execution time with and without probability calculation

Figure 5 shows the results in terms of execution times performed by the two approaches by changing the size of corpus used to answer to the query. The result of this experiment demonstrates that the time needed to return the responses to a request, estimated at 212 s, is negligible. When the size of corpus is 2000, the necessary execution time set by the certain approach is 164 s. In all cases, the time required for indexing our approach is considered negligible compared to standard approach (certain approach) (see Figure 5 We can deduct from these results that our approach gives superior results because it allows us to manage the data uncertainty without consuming more time.

After application of the 2 models on the 3 datasets (KDD, WWW and Nguyen) (KDD: research papers from the ACM Conference on Knowledge Discovery and Data Mining [11], WWW: research papers from the World Wide Web Conference [11] and Nguyen: research papers from various disciplines [40]), we obtain the results described in the

Table 1. This table presents performance metrics (Precision, Recall and$F_{measure}$) for each model and for each dataset. Figures 6, 7 and 8 show, respectively, performance metrics of each model on the KDD, WWW and Nguyen datasets.

Table 2. Evaluation results of approaches in terms of Precision, Recall and $F_{measure}$.

DATASET	Approach	Precision	Recall	$F_{measure}$
KDD	Our approach ISI	0.395	0.254	0.309
	Standard approach	0.212	0.271	0.237
WWW	Our approach ISI	0.211	0.195	0.233
	Standard approach	0.24	0.187	0.210
Nguyen	Our approach ISI	0.23	0.213	0.238
	Standard approach	0.27	0.204	0.224

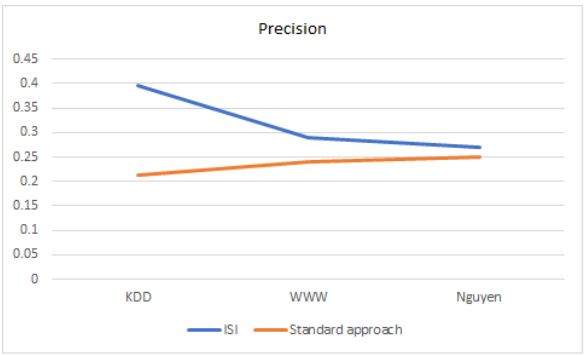

Fig.6. Comparison of assessment results of the evaluation in terms of Precision measurement between the two approaches: Our approach ISI and Standard approach against different standard test databases.

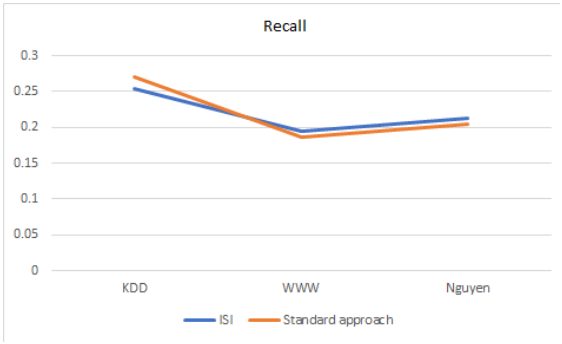

Fig.7. Comparison of assessment results of the evaluation in terms of Recall measurement between the two approaches: Our approach ISI and Standard against different standard test databases approach.

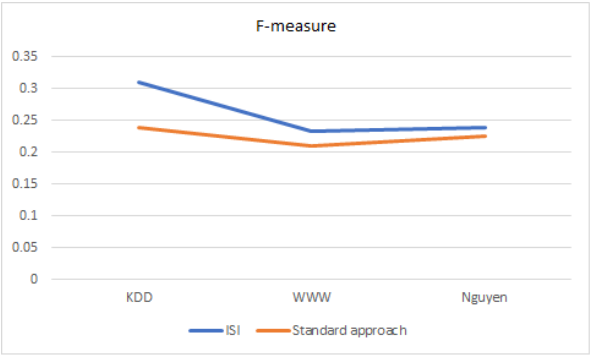

Fig.8. Comparison of assessment results of the evaluation in terms of $F_{measure}$ measurement between the two approaches: Our approach ISI and Standard approach against different standard test databases.

7. Discussion and Limits of the Proposed Approach

Although indexing provides a rigorous foundation in the field of information retrieval for modeling the relationships between documents and terms, the existence of uncertainty in the document retrieval domain remains an unaddressed problem in many other areas such as document indexing. This can be explained by the great complexity of

the concepts of data integration algorithms for big data (such as the case of documentary databases). This complexity limits the evolution of information retrieval systems. Our main idea was to deal with uncertainty and its complexity in IR systems by proposing a model for modeling uncertainty based on a probabilistic approach (which consists of somewhat complex operations). In other words, we propose a new method of syntactic indexing by calculating the certainty values of each term (for each document). Once the index is prepared, our model explores the documents in an intelligent way without modifying the standard indexing structure. We believe that the most important contribution made by our algorithm (ISI) is the simplicity and low complexity that can ensure. The second contribution offered by our algorithm is that it provides precise and automatic rules to calculate the probability of terms. In fact, most information retrieval approaches assume that all data is certain. As shown in the experimental study (see section 6), we have tested our proposed model, and it provides good results. We have shown that our model is more accurate than the standard model in the IKK dataset. However, the response time of our model increased negligibly. This increase can be explained by the uncertainty calculation phase of the terms. Indeed, the response time is a very important factor, among others, for a given information retrieval system. In the real world, users are impatient and don't want to waste time waiting for a response from a given IR system. Moreover, our model can be easily integrated into an information retrieval system. In fact, our ultimate goal is to design an efficient and practical information retrieval system.

The experimental results show the differences, from a performance point of view, between the approach we have proposed and the other approaches. These results also show the factors that influence the quality of these results. The analysis of these results as well as the comparison between the different studied approaches are presented in (Table 1).

Figure 6 illustrates the results obtained for the accuracy rate as well as the difference values from the average accuracy and Figure 7 depicts the recall rate. By analyzing the results obtained from the series of experiments developed for different approaches, we can observe the evolution in terms of performance from one approach to another. The result observed from the term extraction part shows that the performance of our solution is better. The uncertainty calculation phase plays a crucial role in the indexing process because it offers the possibility of keeping as many relevant terms as possible and eliminating those that are not. The proposed approach in this study, compared to others, provides a better result due to determination of values of confidence, which in turn lead to a better representation of documents and terms. The more significant and expressive representation of documents/terms, the more correct the logical deductions on the document basis and terms. According to the results shown in the previous section, we can see the advantages of the proposed approach in terms of important obtained value of 0.395 with KDD database.

According to Figure 7, the approach we have proposed provides a significant value for the recall rate, that is to say more silence (relevant concepts not extracted) at the level of the results obtained. This allowed us to deduce that there are still concepts which have not been extracted then which can represent the document.

Although the cited advantages of the proposed approach, some limits are distinguished. The main objective of the proposed approach is to propose an efficient solution as an indexing approach of Big Data in an uncertain environment. However, this solution despite its good performance in terms of precision, recall and $\boldsymbol{F_{measure}}$ does not automatically manage the uncertainty when it appears when adding new data. This limit penalizes the great performance of our solution and does not allow us to consider it as a perfect solution.

8. Conclusion and Future Work

In recent years, the Web has undergone a major transformation due to great technological development. Indeed, in a context such as the Internet, it is increasingly recognized that data is subject to values of uncertainty, while requiring more sophisticated management techniques. A recent trend is the use of the web as a trusted medium for publishing and sharing data. However, since the Internet has grown exponentially with the increase in the number of sites, data management is becoming a major issue in the information technology industry. Uncertainty and incompleteness are two common characteristics of the information we process in our daily lives. Most areas of informational research do not address this uncertainty. In this paper, we will be interested in the problem of uncertainty at the level of the indexing phase.

Indexing plays a crucial role in retrieving information. In the literature, there are different types of indexing, but in all cases, the principle of the phenomenon is to transform the data sources into electronic data. The purpose of the current study is to propose approaches to deal with uncertain syntactic indexing on textual sources. In this article, we suggested an approach for indexing uncertain sources and sending the user an uncertain response. We proposed two main contributions which are summarized in two main points: (i) Suggestion of an approach to determine the degree of uncertainty of the words that exist in a document and (ii) the integration of this uncertain model of data processing into syntactic indexing.

Our work outlook can be articulated around three directions. The first is to carry out a new, more in-depth comparative study between the main approaches studied, in the literature, and our approach in order to give practitioners and academics more knowledge about indexing big data in an uncertain environment. The second direction consists in extending the research in progress, we intend to integrate the lightweight description logic (DL-lite) which is considered to be one of the most important logics specially dedicated to applications that process large volumes of data. Indeed, the management of inconsistency issues, in order to effectively query inconsistent DL-Lite knowledge bases, is a topical

issue. These inconsistencies in the hierarchical knowledge bases are due to assertions (ABoxes) coming from several sources and therefore are of different levels of reliability. Thus, the addition of new axioms represents a main factor that causes inconsistency in knowledge bases.

References

[1] M-F. Bruandet, J-P. Chevallet, and F. Paradis, « Construction de thesaurus dans le système de recherche d'information IOTA : application a l'extraction de la terminologie ».1eres Journees Scientifiques et Techniques du Reseau Francophone de l'Ingerierie de la Langue de l'AUPELF-URF, Avignon, France. pp. 537–544, 1997.

[2] P. Bosc, and O. Pivert, "About Possibilistic Queries and Their Evaluation", *IEEE Transactions on Fuzzy Systems (TFS)*. Vol.15 no.3, pp.439–452. 2007.

[3] P. Bosc, and O. Pivert, "About projection-selection-join queries addressed to possibilistic relational databases", *IEEE Transactions on Fuzzy Systems (TFS)*. Vol.13 no. 1, pp. 124–139, 2005.

[4] C. Tambellini, "An information retrieval system adapted to uncertain data: adaptation of language model". https://tel.archives-ouvertes.fr/tel-00202702. 2007.Leo Breiman. Random forests.Machine Learning, 45:5–32, 2001.1936.

[5] Z. Lv, X. Li, H. Lv, W. Xiu, "BIM Big Data Storage in WebVRGIS", *IEEE Transactions on Industrial Informatics*. Vol.16 no. 4, pp. 2566 – 2573, 2019.

[6] P. Bosc, N. Lietard, and O. Pivert, "About Inclusion-Based Generalized Yes/No Queries in a Possibilistic Database Context". *ISMIS*. pp.284–289, 2006.

[7] K. Benouaret, D. Benslimane, A. Hadjali, M. Barhamgi, Z. Maamar, and Q. Z. Sheng, "Web Service Compositions with Fuzzy Preferences: A Graded Dominance Relationship-Based Approach", *ACM Transactions on Internet Technology*, Vol. 13 no.4, pp. 1–33, 2014.

[8] Q. He, H. Wang, F. Zhuang, T. Shang, and Z. Shi, "Parallel sampling from big data with uncertainty distribution. Fuzzy Sets and Systems, Vol.25 no. 8, pp. 117–133, 2015.

[9] A. Berko, and V. Alieksieiev, "A Method to Solve Uncertainty Problem for Big Data Sources". 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), pp. 32-37, 2018.

[10] G. Viswanath. and P.V. Krishna, "Hybrid encryption framework for securing big data storage in multi-cloud environment", *Evolutionary Intelligence*. Vol.14: pp. 691–698, 2021.

[11] W. Xizhao, and H. Yulin, "Learning from Uncertainty for Big Data: Future Analytical Challenges and Strategies", *IEEE Systems, Man, and Cybernetics Magazine*, Vol.2 no. 2, pp. 26-31, 2016.

[12] S. Sadeghfam, S. Sadeghfam, A. Nadiri, and K. Ghodsi, "Next Stages in Aquifer Vulnerability Studies by Integrating Risk Indexing with Understanding Uncertainties by using Generalised Likelihood Uncertainty Estimation", *Exposure and Health*. Vol.13, pp.1-15, 2021.

[13] C-C, Lai, H-Y. Lin, and C-M. Liu, "Highly Efficient Indexing Scheme for k-Dominant Skyline Processing over Uncertain Data Streams. *The 30th Wireless and Optical Communications Conference (WOCC 2021)*, 2021.

[14] B. Kao, S. Lee, F. Lee, D. Cheung, and W-S. Ho, "Clustering Uncertain Data Using Voronoi Diagrams and R-Tree Index", *IEEE Transactions on Knowledge and Data Engineering*, vol. 22 no. 9, pp. 1219-1233, 2010.

[15] R. Zhu, B. Wang, and G. Wang, "Indexing Uncertain Data for Supporting Range Queries", *Web-Age Information Management, Springer International Publishing*, 72–83. 2014.

[16] C. Charu, "Aggarwal and Philip S. Yu. On Indexing High Dimensional Data with Uncertainty, 2008.

[17] S. Singh, and C. Mayfield, S. Prabhakar, R. Shah, and C. Hambrusch, "Indexing Uncertain Categorical Data". Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007,

[18] V. Almeida, and R. Güting, "Supporting uncertainty in moving objects in network databases", 13th ACM International Workshop on Geographic Information Systems, ACM-GIS, pp. 31-40. 2005.

[19] D. Veronika, "Indexing Uncertainty: The Case of Turn-Final Or". *Research on Language and Social Interaction*. Routledge. Vol.48 no.3, pp.301-318, 2015.

[20] R. Li, B. Bhanu, Chinya Ravishankar, M. Kurth, J. Ni, "Uncertain spatial data handling: Modeling, indexing and query", *Computers & Geosciences*, Vol. 33, Issue 1, pp. 42-61, ISSN 0098-3004, 2007.

[21] L. M. Mohammed, I. Hamidah, M. Nor Fazlida, Y. Razali,"An Indexed Non-Probability Skyline Query Processing Framework for Uncertain Data", *International Conference on Advanced Machine Learning Technologies and Applications Springer Singapore*, w

[22] Z. Sun, X. Huang, J. Xu, and F. Bonchi, "Efficient Probabilistic Truss Indexing on Uncertain Graphs", *In Proceedings of the Web Conference 2021 (WWW '21). Association for Computing Machinery, New York, NY, USA*, pp.354–366. DOI: https://doi.org/10.1145/3442381.3449976.2021.

[23] M-N. Omri. "Relevance Feedback for Goal's Extraction from Fuzzy Semantic Networks", *Asian Journal of Information Technology (AJIT)*, Vol. 3 no. 6, pp.434-440, 2004.

[24] K. Garrouch, M-N. Omri, and A. Kouzana, "A New Information Retrieval Model Based on Possibilistic Bayesian Networks", *Journal of Information Systems Management*, vol. 2 no.2, pp. 79-88, 2012.

[25] R. Boughamoura, M-N, Omri, and H. Youssef, "A Fuzzy Approach for Pertinent Information Extraction from Web Resources", *International Journal of Computational Science*, vol. 1 no.1, pp.13-30, 2007.

[26] M-N, Omri, "Possibilistic Pertinence Feedback and Semantic Networks for Goal's Extraction", *Asian Journal of Information Technology (AJIT)*, vol. 3 no. 4, pp.258-265, 2004.

[27] Aaron Zimba, Victoria Chama," Cyber Attacks in Cloud Computing: Modelling Multi-stage Attacks using Probability Density Curves", *International Journal of Computer Network and Information Security*, Vol.10, No.3, pp.25-36, 2018.

[28] H-L. Truong, and S. Dustdar, "On analyzing and specifying concerns for data as a service", *IEEE Asia-Pacific Conference on*

Services Computing (APSCC), pp. 87–94, 2009.
[29] A-L. Lemos, F. Daniel, and B. Benatallah, "Web Service Composition: A Survey of Techniques and Tools", *ACM Computing Surveys (CSUR)*. Vol. 48 no. 3, pp.33, 2016.
[30] P. Bosc, and H. Prade, "*An introduction to the fuzzy set and possibility theory-based treatment of soft queries and uncertain or imprecise databases*". SPRINGER. 1994.
[31] P. Smets, "*Imperfect information: Imprecision and uncertainty. In Uncertainty Management in Information Systems*", Kluwer Academic Publishers. 1996.
[32] H. Fukuda, and T-W. Chou, « A probabilistic theory of the strength of short-fibre composites with variable fibre length and orientation", *Journal of Materials Science*, vol.17, pp.1003–1011, 1982.
[33] A. A. Borovkov, "Limit Theorems on the Distributions of Maxima of Sums of Bounded Lattice Random Variables. I. *Theory of Probability & Its Applications*, Vol. 5 no. 2, 1960.
[34] J. Bendler, S. Wagner, T. Brandt, and D. Neumann, "Taming Uncertainty in Big Data - Evidence from Social Media in Urban Areas, *Business & Information Systems Engineering*, Vol. 6 no. 5, pp. 279–288, 2014.
[35] K. Garrouch, and M-N. Omri. "Fuzzy Networks based Information Retrieval Model", *International Journal of Computer Information Systems and Industrial Management Applications*, Vol. 8, 2016.
[36] K. Garrouch, and M-N. Omri, "Possibilistic Network based Information Retrieval Model", *The International Conference on Intelligent Systems Design and Applications (ISDA)*, 2015.
[37] A. R. Pathak. M. Pandey. And S. Rautaray, "Adaptive Model for Dynamic and Temporal Topic Modeling from Big Data using Deep Learning Architecture", *I. J. Computer Network and Information Security*, vol. 6, pp.13-27, 2016.
[38] A. Vasilakopoulos, and V. Kantere, "Efficient Query Computing for Uncertain Possibilistic Databases with Provenance", 3rd Workshop on the Theory and Practice of Provenance (TaPP), 2011.
[39] A. Malki, D. Benslimane, S-M. Benslimane, M. Barhamgi, M. Malki, P. Ghodous, and K. Drira, "Data Services with uncertain and correlated semantics", *World Wide Web*, vol.19, pp.157–175, 2016.
[40] A. Malki, M. Barhamgi, S-M. Benslimane, D. Benslimane, and M. Malki, "Composing Data Services with Uncertain Semantics", *IEEE Transactions on Knowledge and Data Engineering*, vol. 7 no. 4, 936 – 949, 2015.
[41] B. Li, and E. Gaussier, "Modèles de langue pour la recherche d'information'', *Document numérique*, Vol. 16, pp. 11-30, 2013.
[42] M. E. Maron, and J. L. Kuhns, "On Relevance, Probabilistic Indexing and Information Retrieval", *Journal of the ACM*, Vol. 7 no. 3, 1960.
[43] J. M. Ponte, and W. B. Croft, "A Language Modeling Approach to Information Retrieval. SIGIR'98", *Proceedings of the 21st Annual International (ACM) Conference on Research and Development in Information Retrieval*, August 24-28, 1998, Melbourne, Australia. 275–281,1998.
[44] A. Panwar, A. Jain, M. Kumar. A Novel Probability based Approach for Optimized Prefetching. *I.J. Information Engineering and Electronic Business*, 5, 60-67, 2016.

Authors' Profiles

Asma Omri received his Ph.D. in computer science from the University of Claude Bernard Lyon1 (France) in 2018. Since September 2018, she is a lecturer in computer science at the University of Paris-Est Creteil (France). Asma is also a member of MARS (Modeling of Automated Reasoning Systems) Research Laboratory and LACL (Algorithmic, Complexity and Logic Laboratory). Her areas of research include uncertainty, indexing, information retrieval, web, web service, among others.

Mohamed Nazih Omri received his Ph.D. in Computer Science from University of Jussieu, Paris, France, in 1994. He is a professor in computer science at the University of Sousse, Tunisia. From January 2011, he is a member of MARS (Modeling of Automated Reasoning Systems) Research Laboratory. His group conducts research on Information Retrieval, Data Base, Knowledge Base, and Web Services. He supervised more than 20 Ph.D. and Msc students in different fields of computer science. He is a reviewer of many international journals such as Information Fusion journal, Psihologija Journal, and many International Conferences such as AMIA, ICNC-FSKD, AMAI, SOMeT, etc.

How to cite this paper: Asma Omri, Mohamed Nazih Omri, "Towards an Efficient Big Data Indexing Approach under an Uncertain Environment", International Journal of Intelligent Systems and Applications(IJISA), Vol.14, No.2, pp.1-13, 2022. DOI: 10.5815/ijisa.2022.02.01