I.J. Information Technology and Computer Science, **2016, 11, 16-25**
Published Online November 2016 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2016.11.03

Relevant XML Documents - Approach Based on Vectors and Weight Calculation of Terms

Abdeslem DENNAI
University of BECHAR, ALGERIA
E-mail: De_selam@yahoo.fr

Mohammed Yacine DENNAI
University of BECHAR, ALGERIA
E-mail: Med.Yacine.Dennai@gmail.com

Sidi Mohammed BENSLIMANE
LabRI Laboratory Higher School of Computer, SIDI BEL ABBES, ALGERIA
E-mail: S.benslimane@esi-sba.dz

***Abstract*—**Three classes of documents, based on their data, circulate in the web: Unstructured documents (.Doc, .html, .pdf ...), semi-structured documents (.xml, .Owl ...) and structured documents (Tables database for example). A semi-structured document is organized around predefined tags or defined by its author.

However, many studies use a document classification by taking into account their textual content and underestimate their structure. We attempt in this paper to propose a representation of these semi-structured web documents based on weighted vectors allowing exploiting their content for a possible treatment. The weight of terms is calculated using: The normal frequency for a document, TF-IDF (Term Frequency - Inverse Document Frequency) and logic (Boolean) frequency for a set of documents. To assess and demonstrate the relevance of our proposed approach, we will realize several experiments on different corpus.

***Index Terms*—**Semi-structured web document, term weighting, term frequency, TF-IDF and logic frequency.

I. Introduction

XML documents are semi-structured web documents and may be among the components of a web application, their exploitation was regarded as a subject of research strongly advocated in the field of knowledge engineering of the web. The domain that fuses other areas namely: Searching for information (Knowledge) in the web, extraction, acquisition and representation of this information.

These documents have become the means of structuring web data, the most important. With the availability of parsing tools (Such as: SAX, DOM, etc.), you can search and retrieve information relevant to a possible treatment.

The exploitation of web documents as poses many problems for researchers in this domain, they find it difficult to search, retrieve, reliably, information from these documents. To this end, the need to search other representation of these documents is quite crucial, a representation which must be easily interpreted by humans and machines including knowledge engineering programs.

The objective of this paper is to propose a representation of semi-structured web documents approach based weight calculation terms extracted from these documents using the calculation frequencies and subsequently trigger a classification process. This proposed approach goes through four phases: (i). Extraction and structuring: **(i). Extraction and structuring**: This phase allows to extract useful information from XML documents based on the tags, and that after a structuring operation of these documents, **(ii). Calculating the weight of terms**: The information, the result of the previous phase, represents the candidate elements for the current phase where three frequencies are calculated for each term: The normal frequency, TF-IDF frequency and the logic frequency, **(iii). Vector representation**: We represents the extracted terms with their frequencies into vectors, **(iv). Classification**: These vector representations are used to determine the relevant terms and the relevant documents, which will launch a classification process of the documents.

The remainder of paper is organized as follows: Section 2 defines the semi-structured data with examples completed with a recap. Section 3 gives some formulas for calculating the frequency of a term in one or several documents. Section 4 describes some related work. In section 5, we present our contribution with a set of experiments on XML documents. Finally, section 6 concludes with some future perspectives.

II. Semi Structured Data

Semi-structured data are data that are not raw data, or data entered in a conventional data base system. They are

structured data, but they are not organized in a rational model, such as a table or graph-based objects. Many data found on the web can be described as semi-structured. The integration of data allows including the use of semi-structured data [1].

A. Examples of Semi Structured Web Documents

- **XML (eXtensible Markup Language)**

The XML is a computer markup language derived from SGML. This syntax is called "extensible" because it defines different namespaces, that is to say languages with each vocabulary and grammar, as XHTML (eXtensible HyperText Markup Language), XSLT (eXtensible Stylesheet Language Transformations) , RSS, SVG (Scalable Vector Graphics) ... It is recognizable by its use angle brackets (< >) framing tags [2]. The initial objective is to facilitate the automated exchange of complex content (trees, rich text ...) between heterogeneous information systems (interoperability). With its associated tools and languages, an XML application meets certain principles:

- The structure of an XML document is defined and validated by a schema;
- An XML document is fully transformable into another XML document.

The XML standard as such should be seen as a tool to define a language (then said that this is a structuring and markup language or simply a meta-language) for creating structured documents using tags. This meta-language that favored the expression of standard specifications and descriptive standards, such as RDF (Resource Description Framework), DC (Dublin Core), LOM (Learning Object Metadata) or MPEG-7 (Motion Picture Expert Group 7) ... may offer the possibility to create documents that can be treated as an intrinsic database. They are self-descriptive, scalable and above all convertible into several other formats: HTML, XML, PDF (Portable Document Format), RTF (Rich Text File) etc. using style sheets, defined themselves by a language XML called XSL (eXtensible Stylsheet Language). In addition, these documents may conform to structures, themselves based on XML in two existing recommendations that are DTD (Document Type Definition) and XML Schema.

Below is an example of an XML document:

```
<?xml version="1.0" encoding="UTF-8"?>
<xsl:transform  version="1.0"
xmlns:xsl="http://www.w3.org/1999/XSL/
Transform"
xmlns="http://www.w3.org/1999/xhtml"
xmlns:tei="http://www.tei-
c.org/ns/1.0"  >
<xsl:template match="tei:abbr">
<abbr>
<xsl:apply-templates/>
</abbr>
</xsl:template>
</xsl:transform>
```

- **JSON (JavaScript Object Notation)**

JSON is a text data format derived from the notation of JavaScript objects. It is used to represent structured information as XML allows for example. Created by Douglas Rockford between 2002 and 2005, described in RFC 7159 (Request For Comments) of the IETF (Internet Engineering Task Force) [3].

A JSON document serves to represent information along with labels allowing interpreting the various elements, with no restrictions on the number of them.

A JSON document only includes two types of structural elements:

- Sets of pairs name / value
- The ordered lists of values.

These same elements are three types of data:

- Objects,
- Tables,
- Generic values array type, object, Boolean, number, string or zero.

Below is an example of a JSON document: [3]

```
{
"Menu": {
        "Id": "file",
        "Value": "File",
        "Popup": {
            "menuitem": [
{"value":       "New",       "onclick":
"CreateNewDoc()" },
{"Value":       "Open",       "onclick":
"OpenDoc()" },
{"value":      "Close",      "onclick":
"CloseDoc()" }
]
        }
    }
}
```

- **Document Type Definition (DTD)**

The DTD is a document that describes an SGML or XML document template [4].

The model is described as a grammar class documents: grammar, because it describes the position of terms with each other, class because it forms a generalization of a particular area, and the document because can form a complete text.

A DTD describes the documents at two levels: The logical structure, which can be likened to the abstract syntax and physical structure, which can be likened to the concrete syntax.

At the logical structure, a DTD indicates item names that can appear and their contents, that is to say, the sub-

elements and attributes. Besides the attributes, content is specified by indicating the name, sequence and number of times that the sub-elements. The set is the definition of hierarchies' valid elements and text. However, DTDs do not allow ask constraints on the value of the text as "the content of the element X is a decimal integer" or "in the element Y, all white sequences are equivalent to a single space. "Define what is valid is the role of" schemas "as XML Schema, Relax NG (Regular Language for XML Next Generation) and Schematron but these are preferentially expressed in XML syntax while DTDs have a specific syntax. Only DTD is part of the W3C XML, and only possible to validate an XML document from the perspective of this recommendation.

The DTD of a document may be written in and outside of this document. The final DTD is a combination of the two [4].

At the level of physical structure, a DTD can also define general entities. They have one of the following roles: A reference to an external document fragment, typically another file. An abbreviation for repetitive text fragment. For this use, the definition is rather in the internal subset. A synonym for character references by name rather than by a digital code [4].

- **XML Schema**

XML Schema published as a W3C recommendation in May 2001 is an XML document format description language for defining the structure and the content type of an XML document. This definition allows in particular verifying the validity of this document. It is possible to describe an organization of vocabularies of different origins, by the use of namespaces. It is possible to combine the patterns themselves, and express a combination for the content document, such as someone who would talk about geography and sociology in the same text [5].

It is also possible, after validation, to know what rule, particular information was tested: This is the post-schema validation set, or PSVI (Post-Schema-Validation Infoset).

A definition consists of one or more XML documents, usually called XML Schema Definition (or XSD). An instance of an XML Schema is a little equivalent to a document type definition (DTD). However, XML Schema take several differences with the DTD: It can be used to define valid domains for the value of a field, while this is not possible in a DTD; however, it does not define entities; XML Schema is itself an XML document, whereas DTDs are SGML documents.

This XML document content description language is itself defined by a schema, the definition of tags define themselves (This is an example of recursive definition).

The W3C Recommendation 1.0 consists of a presentation document (Non-Normative), a document specifying how to define the structure, and a document specifying how to define the data. The W3C is currently working on version 1.1, which aims to define the schema version notions and constraints depending on the presence of particular value [5].

B. Recap

Table 1. Some examples of semi-structured web documents.

Semi Structured Web Document	Caracteristics
XML	- Extensible Markup Language (predefined markup or to define). - Set different namespaces.
JSON	- Textual data format, derived from the notation of JavaScript objects, has the function to represent information along with labels allowing interpreting the various elements.
DTD	- Document to describe an SGML or XML document template. - Model is described as a class grammar of documents.
XML Schema	- Description Language of XML document format. - Set the structure and the content type of an XML document.

III. Weight Calculations of Terms

The weight of each term in a document can be obtained in different ways: Normal Frequency, TF-IDF and Boolean frequency.

A. Normal Frequency

For term frequency, the weight of a term is obtained by counting occurrences of the term in the document.

B. TF-IDF

A popular and effective technique for associating a weight with terms consists in measuring their rarity, as defined by the measure ***TF-IDF*** *(**T**erm **F**requency - **I**nverse **D**ocument **F**requency)*. The measure of the rarity of a term is calculated as follows: [6]

Suppose a set of documents ***D*** and the frequency of a term ***i*** in a document ***j***. The measurement of the rarity of the term will be determined by its inverse frequency in the document set ***D*** (***IDF***).

The following equation represents the calculation:

$$IDF(t,D) = log\ (\frac{|D|}{|\{d \epsilon D: t \epsilon D\}|}) \quad (1)$$

$IDF(t,D)$: ***IDF*** value of term ***t***.
D: Number of documents in set ***D***.
$d \in D: t \in d$: Number of documents of set ***D*** contained the term ***t***.

For example, if a term is found in all the documents, then the ***IDF*** value of the term will be equal to ***log (1)*** so ***0***. Conversely, if a term is found only in a single document, its value ***IDF*** will be equal to ***log (| D |)***. This

produces the desired effect to lessen the value of words that have no power of discrimination between the documents of the set ***D***.

For the value ***TF***, just count the frequency of a term ***t*** in a document ***d***. Once these calculated values, the following equation is produced which forms the basis of the measure ***TF-IDF***:

$$TF - IDF(t, d, D) = TF(t, d) \times IDF(t, D) \quad (2)$$

The final multiplication gives the ***TF-IDF*** value of a term t from a particular document in a set of documents ***D***. Once these values obtained, determine which documents are most relevant to a given query of user. The use of vector spaces is a way of calculating the correlation between the query and each document of the set ***D***.

C. Boolean Frequency (Logic)

In the Boolean way, if a term exists in a document then the value that corresponds to it is 1, otherwise 0. The Boolean approach is used when each term is equally important and is used only when the documents are small.

IV. Related Works

Our research topic is based on two important aspects: Vector representation approaches of semi-structured web documents and the classification of these documents. The application of this vector representation approaches of semi-structured documents (example: XML) is a highly recommended research topic in the field of knowledge engineering and especially in knowledge representation. Lot of works exist in this field, we will try in this section to cite a few.

S. Chagheri and colleagues in [7] have chosen to not only the classification based on textual content of the document but also taking into consideration the structure. To achieve this, they proposed a weighted vector representation associating a word and a tag or the weight is calculated using the ***TF-IDF*** formulas and ***TF-IEF***. They did an experiment using the classifier SVMlight conducted on Reuters and INEX corpus.

A. M. Vercoustre and his teammates in [8] proposed another classification model. They have taken only the structure or the structure and the content. Their idea was to represent the document by the set of sub paths of the XML tree, the length between n and m (values determined a priori) and they apply to the classification of standard methods such as k-means for example. To evaluate their approach, they conducted experiments on INEX corpus and reports of INRIA.L.

L. Denoyer and colleagues in [9] use the automatic classification of XML documents according to their structural regularities. The main idea is to automatically detect through the structure of the documents, all the sources of information from which they come. This issue finds its meaning in several applications; it can enable viewing by a user of the organization of a body of heterogeneous materials such as Web; it also allows for easier retrieval by selecting the source that a priori most interested user. So they proposed an approach, which aims to show how a generative model for structured documents, based on Bayesian networks, can be used through the Fisher kernel as a model for measuring the similarity between two XML documents. This similarity is assessed through automatic classification task of INEX corpus.

As for our contribution, we have split our process in two. The first: A classification of relevant terms in a document and the second: A classification of documents relevant to a set of documents. In both cases, we have proposed to use three types of frequency calculation as required: Normal Frequency, ***TF-IDF*** and logical frequency. Calculations should be made after an extraction operation terms from XML documents (the details of the approach, discussed in the next section).

V. Our Contribution

Our proposed approach, which is based on calculating the different frequencies of terms extracted from XML documents take into account the structure and content of these documents and includes the following phases:

1. Loading one or more XML documents [10, 11 and 12].
2. Graphical representation of these documents as a tree where nodes are the tags in the XML document [10, 11 and 12].
3. Extracting concepts marked by these tags in XML documents [10, 11 and 12].
4. Calculation of term frequency by using different calculation formulas.
5. Vector representation of these terms and their frequencies.
6. Classification of those documents based on their vector representations.

A. General Approach

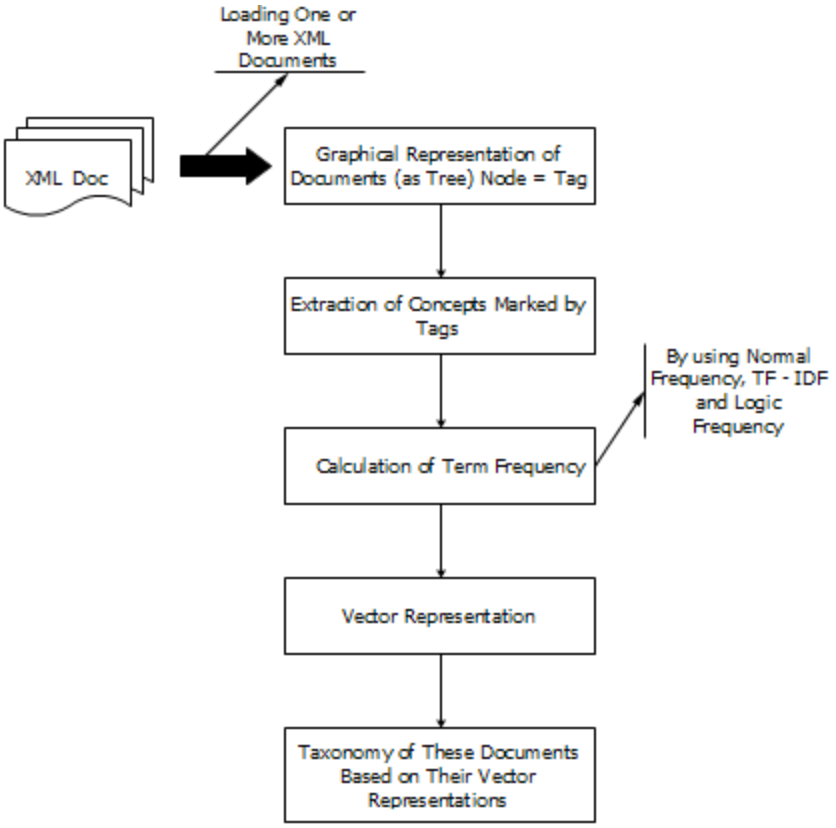

Fig.1. General Approach of Classification.

B. UML Design

Our use of UML is summarized in two essential diagrams: Class diagram and sequence diagram.

- **Class Diagram**

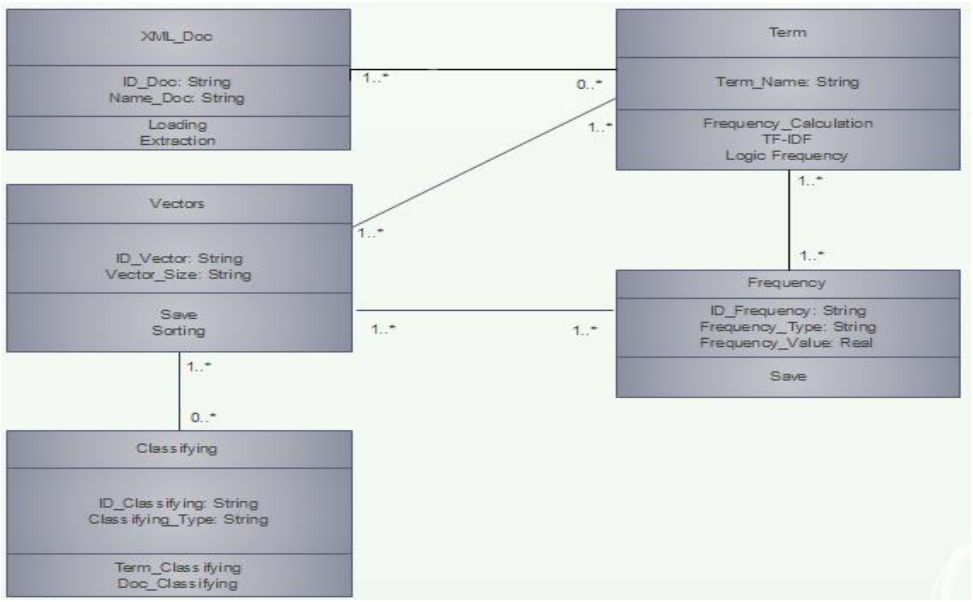

Fig.2. Class Diagram.

- **Sequence Diagram**

For implementation reasons, we are realized four sequence diagrams to explain in detail treatments made by the application.

- **Loading and Extraction.**

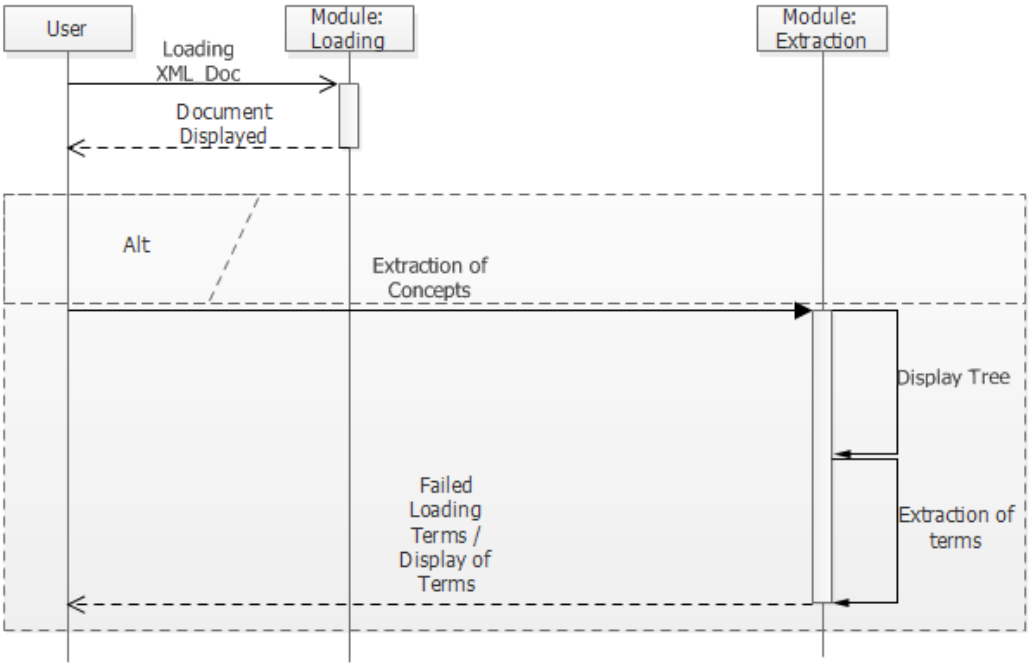

Fig.3. Sequence Diagram (loading and Extraction).

- **Frequencies Calculation**

For one document, it was proposed to apply the frequency calculation using the two formulas: The first simply calculates the number of the word in a document, the second uses the formula ***TF-IDF***.

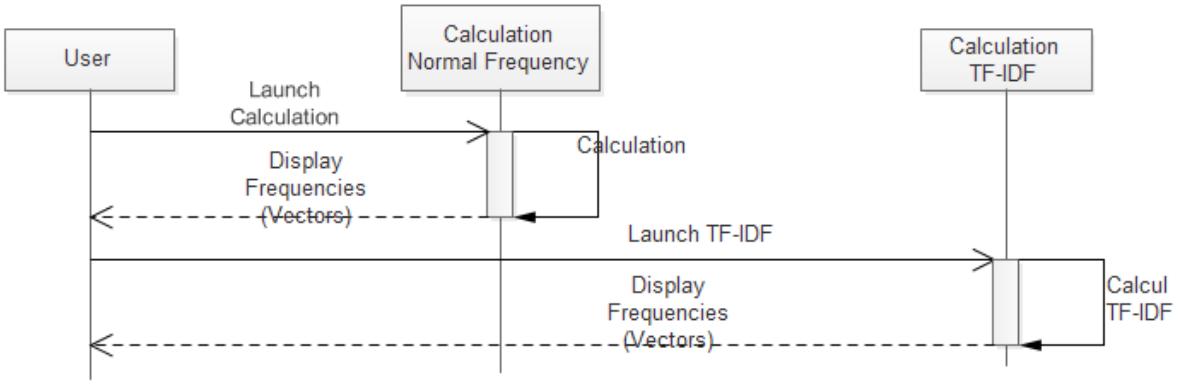

Fig.4. Sequence Diagram (Frequencies Calculation for one Document).

We are proposed to use the calculation of the logic frequency of terms for a set of documents.

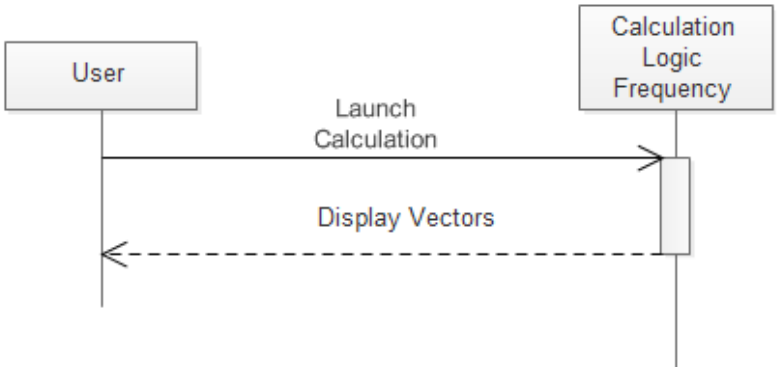

Fig.5. Sequence Diagram (Frequency Calculation of Terms for Several Documents).

- **Classifying of Documents.**

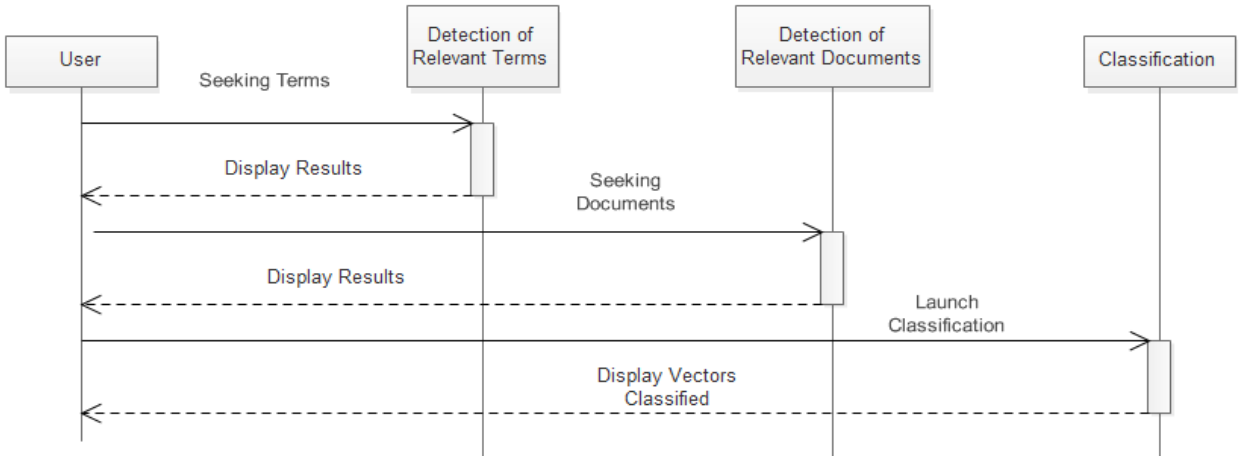

Fig.6. Sequence Diagram (Classifying of Documents).

Note: Classifying operation determines the relevance of terms and the relevance of XML documents.

VI. EXPERIMENTS

A. Example No. 1 (Single Document)

1st XML Document (Doc1)

```
<xml app="DENNAI">
 <item text="Tourist Sites" imageIndex="0" stateIndex="-1">
  <item text="Site" imageIndex="0" stateIndex="-1">
   <item text="Name" imageIndex="0" stateIndex="2"/>
   <item text="Lang." imageIndex="0" stateIndex="2"/>
   <item text="Lang." imageIndex="0" stateIndex="2"/>
   <item text="Place" imageIndex="1" stateIndex="1"/>
  </item>
  <item text="Hotel" imageIndex="1" stateIndex="1">
   <item text="Hotel1" imageIndex="0" stateIndex="2"/>
   <item text="Hotel2" imageIndex="1" stateIndex="2"/>
  </item>
 </item>
</xml>
```

Fig.7. 1er Example of XML Document.

2nd XML Document (Doc2)

```
<xml app="Application">
 <item text="Tourism" imageIndex="0" stateIndex="-1">
  <item text="Sect" imageIndex="0" stateIndex="-1"/>
  <item text="Residences" imageIndex="0" stateIndex="-1">
   <item text="Residence" imageIndex="0" stateIndex="2"/>
   <item text="Residence" imageIndex="0" stateIndex="2"/>
   <item text="Cells" imageIndex="0" stateIndex="2">
    <item text="Cell" imageIndex="1" stateIndex="1"/>
    <item text="Cell" imageIndex="1" stateIndex="1"/>
   </item>
  </item>
  <item text="Regions" imageIndex="0" stateIndex="-1">
   <item text="Region" imageIndex="0" stateIndex="2"/>
   <item text="Region" imageIndex="1" stateIndex="2"/>
  </item>
   <item text="Cells" imageIndex="0" stateIndex="2">
   <item text="Cell" imageIndex="1" stateIndex="2"/>
  </item>
 </item>
</xml>
```

Fig.8. 2nd Example of XML Document.

Table 2. Classifying of Terms based on Frequencies Calculation (1st XML Document).

	Term	Freq.		Term	Rate		Term	Rate
Number of Extracted Concepts = 8 Concepts.	Tourist Sites	1	Ascending Classification (threshold =10 %)	Lang.	22.22	Ascending Classification (threshold =15 %)	Lang.	22.22
	Site	1		Site	11.11			
	Name	1		Name	11.11			
	Lang.	1		Tourist Sites	11.11			
	Place	1		Place	11.11			
	Hotel	1		Hotel	11.11			
	Hotel1	1		Hotel1	11.11			
	Hotel2	1		Hotel2	11.11			

Table 3. Classifying of Terms based on Frequencies Calculation (2nd XML Document).

	Term	Freq.		Term	Rate		Term	Rate
Number of Extracted Concepts = 8 Concepts.	Tourism	1	Classification Croissante (threshold = 10 %)	Cell	23.07	Ascending Classification (threshold = 15 %)	Cell	23.07
	Sect	1		Cells	15.38			
	Residences	1		Residence	15.38			
	Residence	2		Region	15.38			
	Cells	2						
	Cell	3						
	Regions	1						
	Region	2						

3th XML Document (Doc3)

```
<xml app="Application">
 <item text="Tourism" imageIndex="0" stateIndex="-1">
  <item text="Accommodation" imageIndex="0" stateIndex="-1">
   <item text="Hotel" imageIndex="0" stateIndex="2"/>
    <item text="Activity" imageIndex="1" stateIndex="1">
     <item text="Adventure" imageIndex="0" stateIndex="2"/>
     <item text="Relaxation" imageIndex="1" stateIndex="2"/>
     <item text="Sports" imageIndex="1" stateIndex="2"/>
    </item>
  </item>
   <item text="Destination" imageIndex="1" stateIndex="1">
    <item text="Beach" imageIndex="0" stateIndex="2"/>
   </item>
 </item>
</xml>
```

Fig.9. 3th Example of XML Document.

B. 2nd Example (Several Documents)

XML Documents (Doc1, Doc2 et Doc3)

C. Interpreting Tables

In our approach a relevant term is defined in three cases:

1. Its frequency in a document (XML) has a value;
2. The TF-IDF value is measurable;
3. Its frequency in several documents in the same time is important.

To experiment, we were forced to set a threshold having regard to the variation of the frequency of a term values (here the threshold is the value of the frequency of a term relative to the sum of all frequencies of the other terms the same document). This led us to calculate the normal frequency of a term relative to other terms in the same document (see Tables 2, 3 and 4) (Experiences on 3 different XML documents).

We also calculate the TF-IDF frequency of a term, this formula gives the frequency of a term in relation to its rarity in a set of documents (see table no. 5) and that always we give a fixed value for the threshold.

In the same table (Table 5), we determine the logic frequency of a term, which gives the appearance or not of the term in a document. This frequency gives us the relevance of a term relative to several documents. The logic frequency allows also provide relevant documents.

Table 4. Classifying of Terms based on Frequencies Calculation (3th XML Document).

	Term	Freq		Term	Rate		Term	Rate
Number of Extraced Concepts = 9 Concepts.	Tourism	1	Ascending Classification (threshold =10 %)	Tourism	11.11	Ascending Classification (threshold =15 %)	NIL	NIL
	Accom-modation	1		Accom-modation				
	Hotel	1		Hotel				
	Activity	1		Activity				
	Adventure	1		Adventure				
	Relaxation	1		Relaxation				
	Sports	1		Sports				
	Destination	1		Destination				
	Beach	1		Beach				

Table 5. Relevance of Terms and Relevance of Documents based on TF-IDF and Logic Frequency Calculation (Set of three XML Documents: D1, D2 and D3).

	XML DOC	Term	TF-IDF		Term	TF-IDF	Freq. Log.		
							D1	D2	D3
Number of Extracted Concepts = 25 Concepts	Doc1	Tourist Sites	0.60	Ascending Classification (threshold = 0.40)	Tourist Sites	0.60	1	0	0
		Site	0.60		Site	0.60	1	0	0
		Name	0.60		Name	0.60	1	0	0
		Lang.	1.20		Place	0.60	1	0	0
		Place	0.60		Hotel	0.60	1	0	1
		Hotel	0.60		Hotel1	0.60	1	0	0
		Hotel1	0.60		Hotel2	0.60	1	0	0
		Hotel2	0.60		Tourism	0.60	0	1	1
	Doc2	Tourism	0.60		Sect	0.60	0	1	0
		Sect	0.60		Residences	0.60	0	1	0
		Residences	0.60		Regions	0.60	0	1	0
		Residence	1.20		Accom-modation	0.60	0	0	1
		Cells	1.20		Activity	0.60	0	0	1
		Cell	1.80		Adventure	0.60	0	0	1
		Regions	0.60		Relaxation	0.60	0	0	1
		Region	1.20		Sports	1.00	0	0	1
	Doc3	Accom-modation	0.60		Destination	1.00	0	0	1
		Activity	0.60		Beach	1.00	0	0	1
		Adventure	0.60		Lang.	1.20	1	0	0
		Relaxation	0.60		Residence	1.20	0	1	0
		Sports	1.00		Cells	1.20	0	1	0
		Destination	1.00		Region	1.20	0	1	0
		Beach	1.00		Cell	1.80	0	1	0

VII. CONCLUSIONS

In this work, we presented a classification process semi-structured web documents (such as XML Documents). This process, which in part, is based on a vector representation of these documents using the weight calculation terms (frequency). Our process starts with a tree structure of these documents and then extraction of terms, each term undergoes a calculation of it weight by using the calculation of the three frequencies which are: Natural frequency, ***TF-IDF*** frequency and logical frequency. Then all these terms with their frequencies are represented in vectors that will be used for the classification of documents based on the relevance of terms and the relevance of documents. This process allows at the end to have a better representation of documents and facilitation in their exploitations for possible treatment thereafter. The relevance of this process increases, in addition, realizing better extraction of terms and use other frequencies calculations.

The continuity of this research is summarized in three key points, which appear us important in the classification of web documents:

1. Extension of the approach to other types of web Documents, whatever their nature,
2. Use, in addition, other formulas for calculating the frequency of terms
3. Seek other means to determine the relevance of terms and the relevance of documents.

REFERENCES

[1] Moussa L., Amrane H. and Patrick R., "Un modèle de conception d'application Web basé sur XML", ISPS'2001 – Alger, Mai. 2001, RIST Vol. 11 Issue 1, 2001.

[2] W3C Recommendation, "eXtensible Markup Language, 5ème Edition", http : // www.w3.org / TR / 2008/REC-xml-20081126, edited on line Nov. 26 2008, (Consulted June. 2014).

[3] JSON (JavaScript Object Notation), Official WebSite, (Consulted June. 2014).

[4] W3C Recommendation, "Langage de balisage extensible", http://www.w3.org/TR/1998/REC-xml-19980210, Put on line Feb. 10 1998, (Consulted June. 2014).

[5] Hubert K. and Valérie M., "Les web services. Techniques, démarches et outils XML, WSDL, SOAP, UDDI, RosettaNet, UML", Dunod 2003.

[6] Gagnon O., "Indexation de documents web à l'aide d'ontologies", Maitrise en sciences appliquées, Ecole Polytechnique de Montréal, CANADA, 2013.

[7] Chagheri S., Roussey C., Calabretto S. and Dumoulin C, "Classification de documents combinant la structure et le contenu", 2012.

[8] Vercoustre A. M., Fegas M., Lechevallier Y. and Despeyroux T., "Classification de documents XML à partir d'une représentation linéaire des arbres de ces documents", 2006.

[9] Denoyer L., Wisniewski G. and Gallinari P., "Classification automatique de structures arborescentes à l'aide du noyau de Fisher : Application aux documents XML", 6th European Congress on Systems Science, Sep. 19-22, 2005.

[10] Dennai A. and Benslimane S. M., "Information extraction from HTML pages or XML documents by a semantic indexing, using domain ontology", 3rd International Conference on Multimedia Computing and Systems ICMCS'2012, IEEE conference, Tangier, Morocco, 10-12 Mai 2012.

[11] Dennai A. and Benslimane S. M., "Building a Semantic Index from HTML Pages or XML Documents", International Conference on Computing Technology and Information Management, ICCTIM 2014, Dubai, E.A.U, 09- 11 April 2014.

[12] Dennai A. and Benslimane S. M., "Semantic Indexing of Web Documents Based on Domain Ontology",

International Journal of Information Technology and Computer Science (IJITCS), ISSN: 2074 - 9007 (Print), ISSN: 2074 - 9015 (Online), DOI: 10.5815/ijitcs, Published By: MECS Publisher, IJITCS Vol. 7 Issue 2, Jan. 2015.

Authors' Profiles

Abdeslem DENNAI is an Associate Professor at the Computer Science Department of MOHAMMED TAHRI University – BECHAR - ALGERIA. He received his PhD degree in computer science from SIDI BEL ABBES University, in 2015. He also received the diploma of engineering in Computer Science from the University of DJILLALI LIABES – SIDI BEL ABBES - Algeria in 1994. He received the diploma of teaching in Computer Science from the University of BECHAR - Algeria, in 2008. He is currently Member of Research Team 'Service Oriented Computing' at the Evolutionary Engineering and Distributed Information Systems Laboratory, EEDIS. His research interests are in the field of semantic web, web applications and ontology.

Mohammed Yacine DENNAI is a PhD student in first year Computer Science University of BECHAR - ALGERIA. He received his license diploma in Computer Science from the University of BECHAR - Algeria in 2013. He received the master's degree in Computer Science from the University of BECHAR - Algeria, in 2015. His research interests are in the field of semantic web, web applications and ontology.

Sidi Mohamed Benslimane is a full Professor at the Higher School of Computer Science, Sidi Bel-Abbès, Algeria. He received his PhD degree in computer science from Sidi Bel Abbes University in 2007. He also received a M.S. and a technical engineer degree in computer science in 2001 and 1994 respectively from the Computer Science Department of Sidi Bel Abbes University, Algeria. He is currently Head of Higher School of Computer Science, Sidi Bel-Abbès, Algeria. From 2001 to 2015, he was a member of the Evolutionary Engineering and Distributed Information Systems Laboratory, EEDIS. Actually, he heads the Research Team 'Service Oriented Computing' at LabRI-SBA Laboratory. His research interests include, semantic web, service oriented computing, ontology engineering, information and knowledge management, distributed and heterogeneous information systems and context-aware computing.

How to cite this paper: Abdeslem DENNAI, Mohammed Yacine DENNAI, Sidi Mohammed BENSLIMANE,"Relevant XML Documents - Approach Based on Vectors and Weight Calculation of Terms", International Journal of Information Technology and Computer Science(IJITCS), Vol.8, No.11, pp.16-25, 2016. DOI: 10.5815/ijitcs.2016.11.03

