International Journal of Computational Science, Information Technology and Control Engineering
(IJCSITCE) Vol.3, No.3, July 2016

SEARCH OF INFORMATION BASED CONTENT IN SEMI-STRUCTURED DOCUMENTS USING INTERFERENCE WAVE

Larbi GUEZOULI[1] and HassaneESSAFI[2]

[1]LaSTIC, University of Batna,
[2]YottaSwift, France

ABSTRACT

This paper proposes a semi-structured information retrieval model based on a new method for calculation of similarity. We have developed CASISS (Calculation of Similarity of Semi-Structured documents) method to quantify how two given texts are similar. This new method identifies elements of semi-structured documents using elements descriptors. Each semi-structured document is pre-processed before the extraction of a set of descriptors for each element, which characterize the contents of elements.It can be used to increase the accuracy of the information retrieval process by taking into account not only the presence of query terms in the given document but also the topology (position continuity) of these terms.

KEYWORDS

Information Retrieval, Semi-structured information processing, Similarity.

1. INTRODUCTION

Semi-structured data is becoming more and more prevalent. Semi-structured information is entered is the representation of different media like text, video … It facilitates the representation of information in the form of a tree.

There are two types of semi-structured information retrieval: (i) Content Only search (CO) that is based on the textual content of nodes of documents, and (ii) Content And Structure search (CAS) that is based on the textual content and the structure of the nodes of documents[1, 2].

Measuring the similarity between documents or pieces of documents is becoming a hot topic, it can be used in many applications [3-6]. Many techniques are proposed to measure the similarity between two documents. It can be grouped into two classes: similarity based schema and similarity based content. In our knowledge, the approaches considering the contents use only terms as key elements of similarity measure. In our CASISS approach, we exploit both the structure and the content. Furthermore, to detect the paraphrasing resemblance, we take into account in the measure of the similarity the neighbourhood of terms.

The semi-structured information retrieval can be based on the need of the content only or the content and the structure, but generally both Content Only (CO) and Content And Structure (CAS) variants are requested [7, 8].

In semi-structured data, similar concepts are represented using different types, heterogeneous sets are present and object structure is not fully known[1, 2].

For the evaluation of semi-structured information retrieval systems, there is, actually, only one company, called INEX (INitiative for the Evaluation of Xml retrieval), of evaluation of performances of semi-structured information retrieval systems that is based on Recall/Precision measure. INEX is presented like a program uses a collection of semi-structures documents with a set of topics as well as relevance judgments.

In this paper we propose a new model of semi-structured information retrieval. It is based on a new method of calculation of similarity. We aim, with our approach, the Content And Structure variant of the information need.

The main novelties of CASISS model are: (i) the definition of a new concept called interference wave; and (ii) using this wave to compare semi-structured documents by using the context of terms represented by neighborhood.

The rest of paper is organized as follows: Section2 introduces works in relation to our work, section0 describes the background of our solution, and in section4 we define our model. Experimental results obtained from implementing our approach are shown in section 5. Finally, we conclude this paper with some perspectives of our work.

2. STATE OF THE ART

Schema matching is quite old problem[9]. However, with the growing use of XML format as standard for exchanging, matching XML content gained in interest.

In the recent years, researchers study the indexation of nodes. A survey on semi-structured documents mining was presented by Madani et al. [10].They done a comparison between several existed approaches using different comparison criteria like used technique, complexity of algorithm, etc.

Another survey is presented by Leena A Deshpande and R.S. Prasad [11]. They give a brief survey of various data mining techniques and recent research issues for representing semi-structured databases, especially XML.

In this section, some related works on semi-structured information retrieval models will be presented.

One of them is the work of Haïfa Zargayouna presented in his thesis [12] in which she indexes XML documents semantically. Another work of Burke R. et all [1] in which they propose semi-structured information retrieval based on knowledge and they uses their model to develop a FAQ finder tool.

Zhang et al. [13] base their searches on the knowledge of the user, which is not always available.Another work based on algebraic approach, proposed by Ben Aouicha [14]. He proposes an algorithm for the comparison between trees in order to localize sub-trees similar to the tree of the query.

Saikat G. and Chandan K. propose the use of the XML distributed database [15]. They implemented their proposed database on library system. In our case, we can use this genre of database.

Lipczak et al. [16] propose a selective retrieval for categorization of semi-structured web pages. Their approach is practically usable for real-time interaction, but it limits the need for the retrieval of additional information.

An interesting work presented by Xue-Liang Zhang et al. [17] where they present a new method of computing the structure and semantic similarity of XML documents based on extended adjacency matrix (EAM). Their approach calculates the similarity between two semi-structured documents, but we need to compare one document with a database.

Many approaches exist to retrieving information from distributed heterogeneous semi-structured documents like the work of CHOE et all. and Mone [2].They concentrate on problems caused by the distributed environment. Aditya [18]presents in his thesis an approach of flexible retrieval system for semi-structured documents based on the vector space model like works cited in[8, 12]. The flexible approach doesn't process the semantic. Filippo Geraci and Marco Pellegrini[19] devise an alternative way of embedding weights in the data structure, coupled with a non-trivial application of a clustering algorithm based on the furthest point. The notion of semantics is absent.Renaud Delbru et al. work on a project called "SIREn: Efficient semi-structured Information Retrieval for Lucene" [20],they advocate the use of a node indexing scheme for indexing semi-structured data. They base their indexing method on structure of document. The content is not taken into account.

3. BACKGROUND

An information retrieval system includes generally five components: query, indexing scheme, similarity measure, threshold, and collection. Query represents user's information need. Indexing scheme is a representation model of relations between terms and documents. Similarity measure is a function which determines the degree of resemblance between user's query and a document. Threshold is a real number which indicates how we should filter out irrelevant documents.
With our approach, user can express his information need by a semi-structured document and can choose between content only search or content and structure search.

When user issues a query, it will be converted to three parts: Content, structure and links between content and structure. The conversion of given query passes by two steps:

- **Pre-processing**

In this step we extract the contents of nodes, then we extract the graph of the structure of the document and finally we store links between contents and parts of extracted graph.

- **Normalization of contents**

In this step we extract lexical units from contents of nodes. To do this process we'll first delete stop words, then change terms by their roots (stemming or lemmatization).

4.CASISS MODEL

Our proposed model CASISS (Calculation of Similarity of Semi-Structured documents) is different from works cited in section 3. It is based on a new concept called "interference wave". Fig. 1 represents the different steps of proposed model.

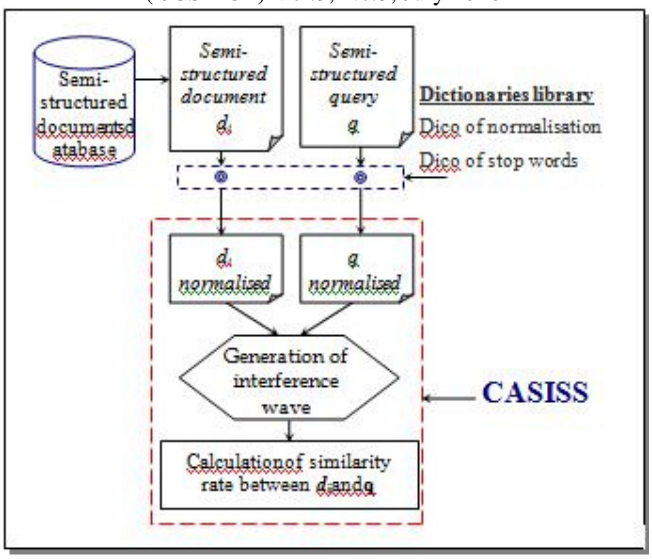

Fig.1 General scheme of CASISS model

After preliminary steps, pre-processing and normalization of contents, we present the indexing scheme by a tree.

Fig. 2 presents the representation scheme of a semi-structured document before the construction of indexing scheme.

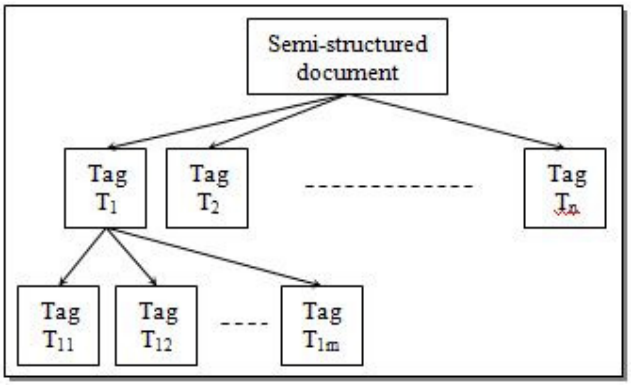

Fig.2 Representation of a semi-structured document by a tree

We use this representation to make the scheme of indexing by extracting the frequency of each term in the content of a node, the list of positions of this term and its neighborhood.

For each term, we need its number of occurrences and their positions to locate it after the end of research process. We need also for each occurrence its left and right neighbor.

Each node will have its fingerprint independently of other nodes called *node-fingerprint*. The set of *nodes-fingerprints* of all nodes of corpus are used to construct the global indexing scheme.
When a query is presented to the system, it passes also by preliminary steps: pre-processing and normalization of contents. Then, the system extracts the set of *nodes-fingerprints* of given query.
After the extraction of all *nodes-fingerprints* of all documents of corpus and query, we generate the interference wave. It's used to calculate the rate of similarity between each semi-structured document of the base and the semi-structured query.

4.1 INTERFERENCE WAVE

The goal is to model the comparison between significant terms of semi-structured documents. The idea is to define a function that associates a value from the set $\{e_0, e_1, e_2\}$ to the triplet (t_i, d_j, q).

We call the graph of this function "interference wave" and we denote γ, as shown in Fig. 3.

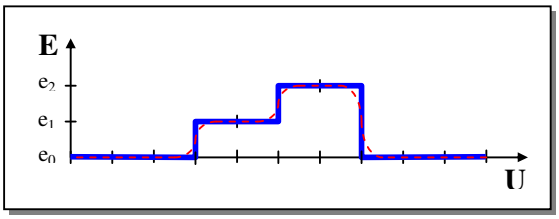

Fig.3 Interference wave γ

The function γ is defined as follows:

γ : $U \rightarrow E$

$u \mapsto \gamma(u) = e$

where: U is the vocabulary set and $e \in E = \{e_0, e_1, e_2\}$.

Let:

- $PQuest_j$ is the content of j^{th} node of semi-structured query.
- $PBase_i$ is the content of i^{th} node of semi-structured document of corpus.
- $path(PQuest_j)$ is the path from the root node to the j^{th} node of semi-structured query. A path is a string composed by the nodes titles separated by points.
- $path(PBase_i)$ is the path from the root node to the i^{th} node of semi-structured document of corpus.

The meaning of the three values of the set E is:

- $\gamma(u) = e_0$ if $PQuest_j$ matched with $PBase_i$ and $path(PQuest_j)$ matched with $path(PBase_i)$.
- $\gamma(u) = e_1$ if $PQuest_j$ matched with $PBase_i$ and $path(PQuest_j)$ is different from $path\ (PBase_i)$.
- $\gamma(u) = e_2$ if $PQuest_j$ is different from $PBase_i$.

with:

- $PQuest_j$ is different from $PBase_i \Rightarrow \left(CASIT(PBase_i, PQuest_j) > threshold\right)$.
- $PQuest_j$ matched with $PBase_i \Rightarrow \left(CASIT(PBase_i, PQuest_j) \leq threshold\right)$.

where CASIT is a similarity measure between two textual documents. This measure is described in our previous publication [21], it takes into account the neighborhood of words. Therefore, instead of comparing two semi-structured documents we use the interference wave to calculate the rate of similarity between these documents. To calculate the rate of similarity we convert the interference wave to two interference vectors.

4.2 INTERFERENCE VECTORS

From the interference wave γ, we establish the two interference vectors V_0 and V_1. The vector V_0 is built by using the sequences of terms $u \in U$ such as $\gamma(u) = e_0$ and V_1 is obtained by using the sequences of terms $u \in U$ such as $\gamma(u) = e_1$.

Let *n-gram* is a sequence of n elements with the same value.

The element of the rank n in the vector V_i contains the number of *n-grams* of level i in the interference wave γ (sequences of n elements of value e_i).

number of 1-grams	number of 2-grams	number of 3-grams	……	number of n-grams

V_i : Interference vector of the level i

In the example of the Fig. 3, the interference vectors are:

0	0	2

V_0 : Interference vector of the level **0**

0	1	0

V_1 : Interference vector of the level **1**

For example:

$V_0[3] = 2 \Leftrightarrow$ there exist **2 3**-grams of level **0** in the interference wave.

$V_1[2] = 1 \Leftrightarrow$ there exist **1 2**-grams of level **1** in the interference wave.

After establishing of interference vectors we can calculate the similarity.

4.3 SIMILARITY

The function ϖ fixes the similarity rate between semi-structured query **Q** and semi-structured document **D** of the corpus. This function ϖ is defined by using interference vectors V_0 and V_1 as follow:

$$\varpi = \frac{2 \times \sum_{j=1}^{n} j \cdot V_0[j] + \sum_{j=1}^{m} j \cdot V_1[j]}{2 \cdot n + m} \times 100 \qquad \text{Eq. (1)}$$

where n is the size of V_0, m is the size of V_1.

The denominator allows normalizing of the similarity rate between 0% and 100%, because the maximal value of the nominator is $2.n+m$.

Next figures present different cases of comparison between two semi-structured documents.
Fig.1 presents the interference wave of comparison between two similar semi-structured documents. The query document contains noises. The final score (calculated using equation Eq. 1) of this case is 76%.

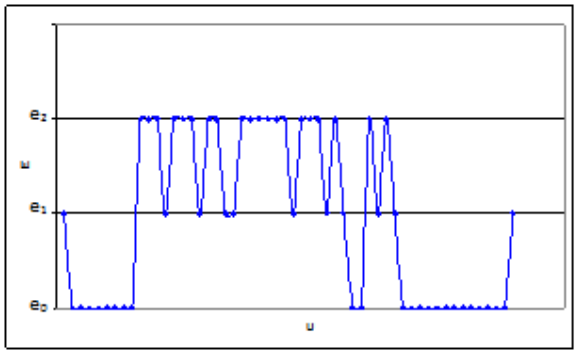

Fig.1Interference wave of the comparison between two similar semi-structured documents with noises

Fig. presents the interference wave of a comparison between two similar semi-structured documents. The final score of this case is 99%.

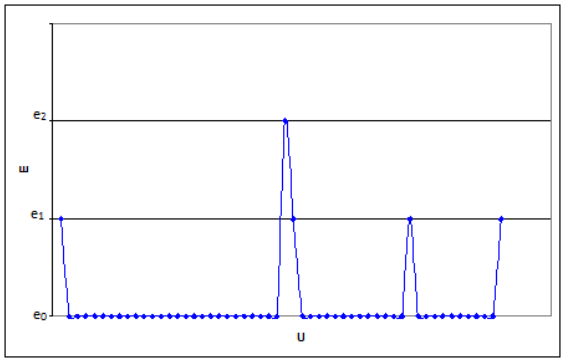

Fig.5 Interference wave of the comparison between two similar semi-structured documents

Fig. presents the interference wave of a comparison between two different semi-structured documents. The final score of this case is 1%.

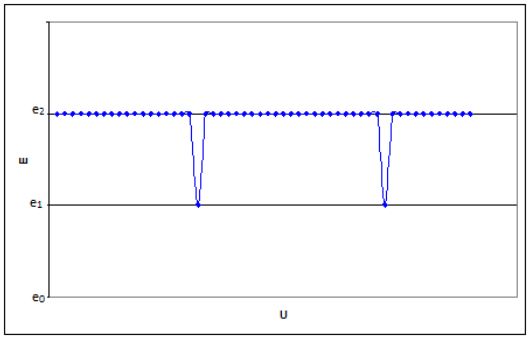

Fig.6 Interference wave of the comparison between two different semi-structured documents

4.4 EXPERIMENTS

To evaluate the performance and reliability of our system, the time was determined as a benchmark.

Experiments have been applied to a database of 11080 XML documents, in French language. Its size is 23.4 Mo with four different disciplines, extracted following an analysis of the 2011 INEX collection[22]. which is characterized by its heterogeneity. These documents are known by the variance of their size, for that, the evaluation is made based on the number of terms in documents.

4.5 TEST OF INDEXING TIME

The graph of the figure bellow shows the change in indexing time based on the number of terms in the XML documents:

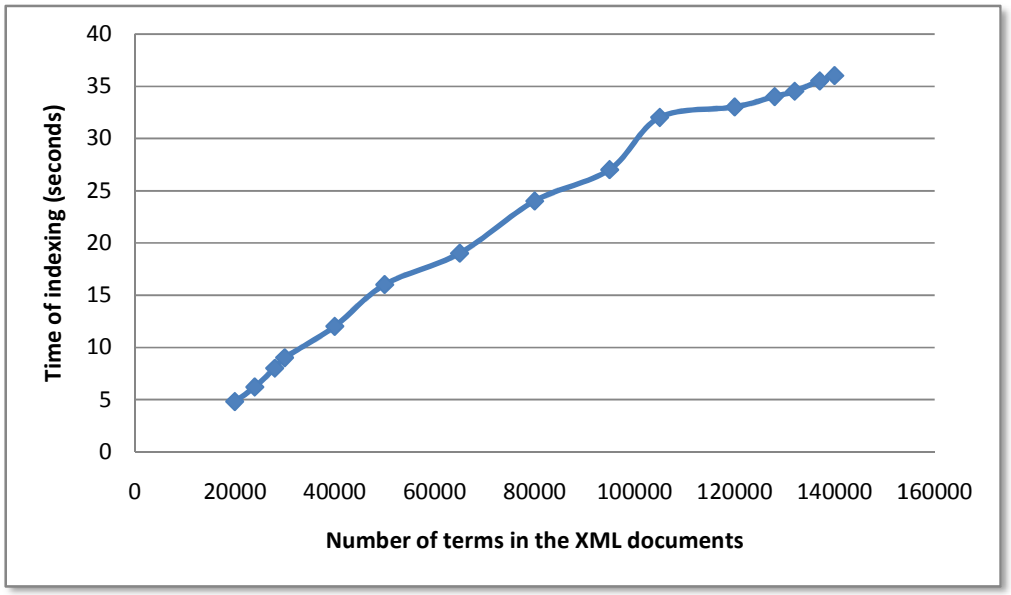

Fig.7. Evolution of the time of indexing according to the number of terms in the XML documents

We observe that the graph mounts linearly, which means that the execution time increases in a linear way (not exponential). This makes the use of large databases feasible (running time over).

4.6 TEST OF RESEARCH TIME

The graph below shows the change in the research time according to the number of terms in XML documents.

We observe that the graph mounts almost linearly, so execution time increases in a linear way which makes the search time over. It's due to the increased size of documents in the database, leading automatically to the increase in the number of terms in XML documents of the database.

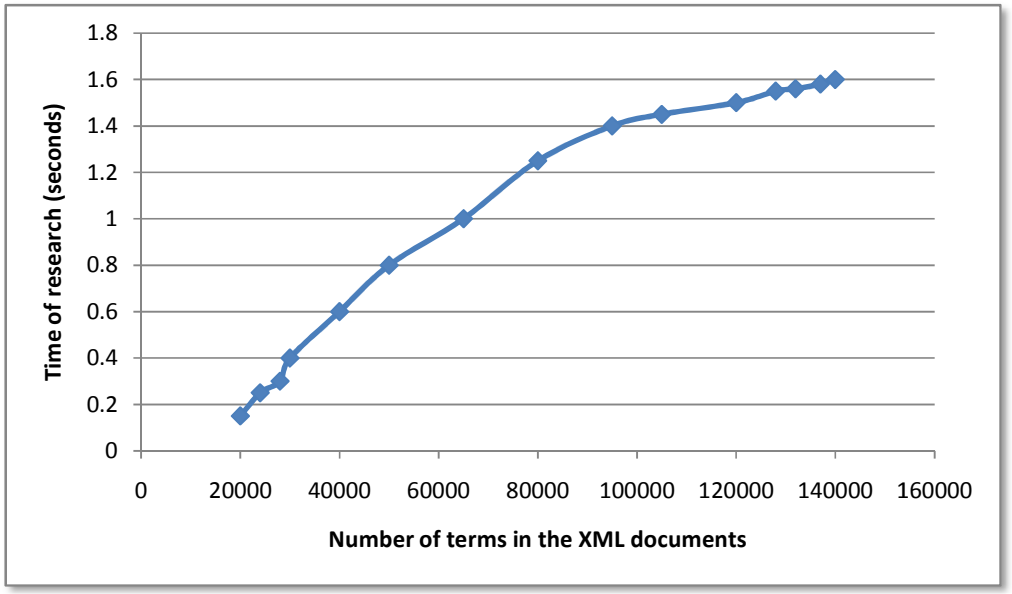

Fig.8 Evolution of the search time according to the number ofter ms in the XML database

In order to measure the capacity of our system to distinguish the XML documents similar to the required XML document, we use the Recall/Precision measure. This measure is defined like performances evaluation, applied to three various requests in order to have various results. These requests treat the cases of homonymy and synonymy as was shown on the Fig Fig.9.

As we know, a large number of returned documents correspond to a high recall rate, but a rather low accuracy rate, while a small number represent the opposite.

We observe on the curve of Fig.9 that with a medium recall rate we have a good accuracy (0.4, 0.78). This means that with a small number of returned documents we get a good accuracy.
On the other side of the curve, we observe that a good recall rate corresponds to a medium accuracy rate (0.8, 0.35). This means that despite the large number of returned documents, the accuracy is acceptable.

This curve shows clearly that our system offers good performance.

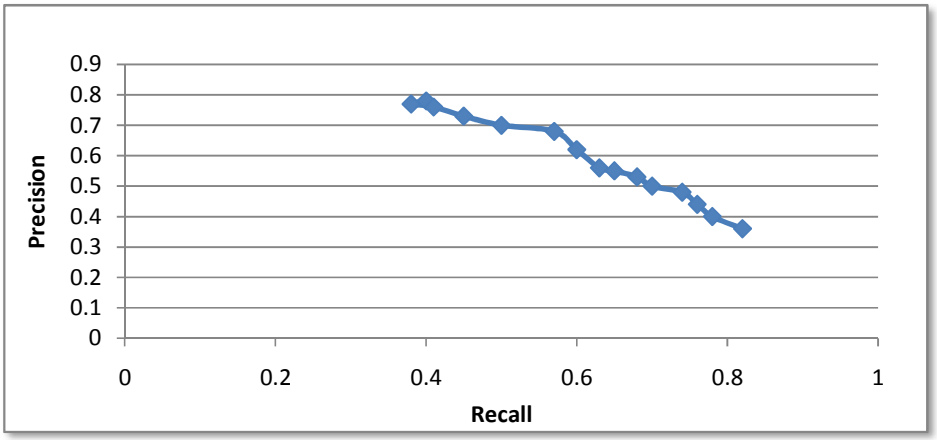

Fig.9 Precision-Recall curve obtained by the mean of results of three requests.

5 CONCLUSION

In this paper we have presented a new similarity measure, dedicated to the semi-structured information retrieval. It uses the interference wave which is generated by applying a similarity function to nodes of XML required document towards those of documents of database. It's based on two types of information: structural and textual.

Structural information is represented by the path (sequence of nodes) leading to textual information. Textual information is represented by the vocabulary associated with its neighborhood. Using the neighborhood, we benefit from the semantics of sentences, and improving the relevancy of the research. The carried system is constituted by the two phases: indexing and search.

This system is tested by using a prepared corpus (extracted from INEX collection) according to the execution time. We had found that the execution time is linear and finite. Then, we evaluated our method using the Recall/Precision graph in order to measure the quality and relevance of answers returned by the system.

REFERENCES

[1] Burke ,R., K. Hammond, and E. Cooper. Knowledge-Based Information Retrieval From Semi-Structured Text. in AAAI Workshop on Internet-based Information Systems. 1996.

[2] Choe, G., et al., Information Retrieval from Distributed Semistructured Documents Using Metadata Interface, in Knowledge Discovery from XML Documents, R. Nayak and M. Zaki, Editors. 2006, Springer Berlin Heidelberg. p. 54-63.

[3] Algergawy, A., et al., XML data clustering: An overview. ACM Comput. Surv., 2011. 43(4): p. 1-41.

[4] Guerrini, G., M. Mesiti, and I. Sanz, An overview of similarity measures for clustering XML documents, in Web Data Management Practices: Emerging Techniques and Technologies, A.V.a.G. Pallis, Editor. 2007, Idea Group Inc.

[5] Sekhar.K, C. and Dhanasree, Extracting TARs from XML for Efficient Query answering. International Journal of Computer Science and Network, 2012. 1(6): p. 40-45.

[6] Zerdazi, A. and M. Lamolle, Computing Path Similarity Relevant to XML Schema Matching, in Proceedings of the OTM Confederated International Workshops and Posters on On the Move to Meaningful Internet Systems: 2008 Workshops: ADI, AWeSoMe, COMBEK, EI2N, IWSSA, MONET, OnTo Content + QSI, ORM, PerSys, RDDS, SEMELS, and SWWS. 2008, Springer-Verlag: Monterrey, Mexico. p. 66-75.

[7] Manning, C.D., P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge University Press ed. 2008.

[8] Salton, G. and M.J. McGill, Introduction to Modern Information Retrieval. 1986, New York, NY, USA: McGraw-Hill, Inc. 400.

[9] Castano, S. and V.d. Antonellis, A Schema Analysis and Reconciliation Tool Environment for Heterogeneous Databases, in Proceedings of the 1999 International Symposium on Database Engineering & Applications. 1999, IEEE Computer Society. p. 53.

[10] Madani, A., O. Boussaid, and D.E. Zegour, Semi-structured Documents Mining: A Review and Comparison. Procedia Computer Science, 2013. 22(0): p. 330-339.

[11] Deshpande, L.A. and R.S. Prasad, Efficient Frequent Pattern Mining Techniques of Semi Structured data: a Survey. International Journal of Advanced Computer Research, 2013. 3(8): p. 177-181.

[12] Haïfa, Z., Indexation sémantique de documents XML. 2005, University of Orsay.

[13] Zhang, L., Y. Zhang, and Q. Xing, Filtering semi-structured documents based on faceted feedback, in Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. 2011, ACM: Beijing, China. p. 645-654.

[14] Ben-Aouicha, M., Une approche algébrique pour la recherche d'information structurée. 2009, University of Paul Sabatier of Toulouse.

[15] Goswami, S. and C. Kundu, XML based advanced distributed database: implemented on library system. International Journal of Information Management, 2013. 33(1): p. 28-31.
[16] Lipczak, M., et al., Selective Retrieval for Categorization of Semi-structured Web Resources, in Advances in Artificial Intelligence, O. Zaïane and S. Zilles, Editors. 2013, Springer Berlin Heidelberg. p. 126-137.
[17] Zhang, X.-L., et al., Novel Method for Measuring Structure and Semantic Similarity of XML Documents Based on Extended Adjacency Matrix. Physics Procedia, 2012. 24, Part B(0): p. 1452-1461.
[18] Mone, A.S., Dynamic Element Retrieval for Semi-Structured Documents. 2007, University Of Minnesota: USA. p. 61.
[19] Geraci, F. and M. Pellegrini. Dynamic user-defined similarity searching in semi-structured text retrieval. in Proceedings of the 3rd international conference on Scalable information systems. 2008. Vico Equense, Italy: ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).
[20] Delbru, R., et al. A node indexing scheme for web entity retrieval. in Proceedings of the 7th international conference on The Semantic Web: research and Applications - Volume Part II. 2010. Heraklion, Greece: Springer-Verlag.
[21] Guezouli, L. and H. Essafi. CASIT: Content Based Identification of Textual Information in a Large Database. in Advanced Information Networking and Applications Workshops (WAINA), 2010 IEEE 24th International Conference on. 2010.
[22] Bellot, P., et al. INEX 2011 Question Answering Track. 2011 [cited 2012; Available from: https://inex.mmci.uni-saarland.de/tracks/qa/2011/.

AUTHORS

Larbi GUEZOULI (b. 1974) received the Bachelor degree in mathematical techniques from the technical high school of Batna, Algeria, in 1992, the engineering degree in computer science from the University of Batna, Algeria, in 1997 and Ph.D. degree in computer science from the University of Paris 7 (Denis Diderot), France, in 2007. He was also a Vice Director of computer science department of the University of Batna in Algeria from 2010 to 2016. Actually, he is a teacher researcher at computer science department on Batna University. His major interests are information retrieval, data mining, and cross language. He is an author of more than 9 conference proceedings papers, 4 journals papers and a patent in his research areas.

Hassane ESSAFI PhD Founder, Chief Scientific Officer and VP of Advanced R&D PhD in Information Sciences, HassaneEssafi is an enthusiastic developer of technologies destined to have maximum impact on industry. After his doctorate, Mr. Essafi joined the Technology Research directorate of the French Atomic Energy Commission (CEA) to facilitate industrialization of his doctoral research into parallel systems for image analysis. The second generation of these products forms an integral part of the onboard vision systems of the Rafale fighter aircraft. He continues his work with the CEA, focusing on new and challenging R & D issues in emerging industries. Dr. Essafi has over 50 publications to his name, as well as 10 patents, and has managed a number of international research projects. Accredited as an expert in Information Technologies by European Union authorities, he is a member of the board of the French Society for Image Recognition and Interpretation.