HOSTED BY
ELSEVIER

Available online at www.sciencedirect.com
ScienceDirect
journal homepage: www.elsevier.com/locate/jides

CAS-based information retrieval in semi-structured documents: CASISS model

Larbi Guezouli[a,*], *Hassane Essafi*[b,c]

[a] *LaSTIC, University of Batna 2, Algeria*
[b] *CEA, France*
[c] *YottaSwift, France*

HIGHLIGHTS

- The precision of the proposed approach is 100%.
- The semantic of words is used in the calculation of the similarity by using neighboring.
- Good performances supplied by the interference wave.
- The execution time is acceptable.

ARTICLE INFO

Article history:
Received 18 March 2016
Received in revised form
2 November 2016
Accepted 6 November 2016
Published online 5 December 2016

Keywords:
CASISS
Information retrieval
Semi-structured information
processing

ABSTRACT

This paper aims to address the assessment the similarity between documents or pieces of documents. For this purpose we have developed CASISS (CAlculation of SImilarity of Semi-Structured documents) method to quantify how two given texts are similar. The method can be employed in wide area of applications including content reuse detection which is a hot and challenging topic. It can be also 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 (Content Only search — CO) but also the topology (position continuity) of these terms (based on Content And Structure Search — CAS). Tracking the origin of the information in social media, copy right management, plagiarism detection, social media mining and monitoring, digital forensic are among other applications require tools such as CASISS to measure, with a high accuracy, the content overlap between two documents.

CASISS identify elements of semi-structured documents using elements descriptors. Each semi-structured document is pre-processed before the extraction of a set of elements descriptors, which characterize the content of the elements.

Peer review under responsibility of Qassim University.

[*] *Corresponding author.*
E-mail address: larbi.guezouli@univ-batna.dz (L. Guezouli).
URL: http://fac-sciences.univ-batna.dz/cs/enseignants/guezouli_larbi_site/index.html (L. Guezouli).

http://dx.doi.org/10.1016/j.jides.2016.11.004

1. Introduction

Semi-structured data is becoming more and more prevalent. Semi-structured information is used in 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 [3–7]. 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, these approaches considering the contents use only terms as key elements of similarity measure. In our proposed approach called CASISS (Calculation of Similarity of Semi-Structured documents), we exploit both the structure and the content. Most of the existing models don't take into account aspects of term's neighborhood in a document.

For example, a term of query with the same neighbourhood in a document gives us an important clue that cannot be achieved by existing models. 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 [8,9]. 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 campaign, 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-structured 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, both the Content Only and the Content And Structure variants of the information need.

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

The rest of paper is organized as follows: Section 2 introduces works in relation to our work, Section 3 describes the background of our solution, and in Section 4 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. 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 indexing of nodes. A survey on semi-structured documents mining was presented by Madani et al. [10]. They have 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 gave 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 Hafa 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 al. 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 [9,12]. 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.