International Journal of Advanced Engineering, Management and Science (IJAEMS)
Peer-Reviewed Journal
ISSN: 2454-1311 | Vol-7, Issue-8; Aug, 2021
Journal Home Page: https://ijaems.com/
Article DOI: https://dx.doi.org/10.22161/ijaems.78.1

Comparative Study on Graph-based Information Retrieval: the Case of XML Document

Imane Belahyane[1], Mouad Mammass[2], Hasna Abioui[2], Assmaa Moutaoukkil[3], Ali Idarrou[2]

IRF-SIC Laboratory, Ibn Zohr University, Agadir Morocco

Received: 11 Jul 2021; Received in revised form: 04 Aug 2021; Accepted: 12 Aug 2021; Available online: 20 Aug 2021

Abstract— *The processing of massive amounts of data has become indispensable especially with the potential proliferation of big data. The volume of information available nowadays makes it difficult for the user to find relevant information in a vast collection of documents. As a result, the exploitation of vast document collections necessitates the implementation of automated technologies that enable appropriate and effective retrieval.*

In this paper, we will examine the state of the art of IR in XML documents. We will also discuss some works that have used graphs to represent documents in the context of IR. In the same vein, the relationships between the components of a graph are the center of our attention.

Keywords— ***Information Retrieval, graph relations, graph-based approach, XML document, XML document retrieval.***

I. INTRODUCTION

In the context of BigData, The continual spread of digital multimedia documents necessitates the updating of the many existing technologies used to exploit the available digital mass. Consequently, locating information is a process of matching the query expressing the user's need with the documents in the documentary database.

In a digital world, where the number of documents grows exponentially, studies have shown that more than 80% of the organizations and businesses' data is in the form of documents, compared to less than 20% in traditional databases. Our research operates within the framework of the Information Retrieval (IR) in semi-structured documents (SSD). Indeed, the SSD has organizational properties that facilitate their analysis.

In the literature, the XML document (eXtensible Markup Language) is a documentary standard, which is generally qualified as a semi-structured document. Indeed, the XML document is quickly becoming the ultimate format for exchanging documents - and, more typically, information. It is currently utilized by ERP providers, middleware editors, database providers, and in e-commerce and libraries. To retrieve information from a document corpus, a representation model that is appropriate for the type of document is required. The more sophisticated the modeling of the documents will be, the more relevant but difficult the comparison of the documents will be[1] . The standard structuring languages such as XML and their derivatives allow serializing a tree structure of a document in the same file. As a result, the DSS must be represented by a graph that measures the many relationships between the components of the XML document and prioritizes the semantic and contextual information conveyed.

The rest of the paper is organized as follows: In section 2, we will describe the principle of Document Information Retrieval, by establishing a state of art on the works that addressed the issue of Information Retrieval in XML documents. In section 3, we will highlight the types of relationships that exist between the components of the graph. In section 4, we will analyze the relevance of the structural aspect, specifically the relations of a graph, as well as the advantages and limitations of the many types of relations that exist between the components of the graph, to determine which would best serve our context.

II. INFORMATION RETRIEVAL IN XML DOCUMENTS

2.1 Introduction

Information Retrieval is a field of computer science that deals with the acquisition, organization, storage, search and selection of information [2], That system can locate precisely documents related to a query expressing a user's need.

Globally, the classic steps of an IRS are the indexing phase, the representation phase, and the comparison phase. During the indexing phase, the system unifies the coding of the documents in the document collection and organizes the collection. At the level of the document representation phase, documents are represented by a model that synthesizes them as much as possible. In the comparison phase, the system compares the query to all of the documents in the collection. It does so by providing a matching mechanism between the user query and the documents, or more particularly between the query representative and the document representatives, the system compares the query to all of the documents in the collection.

IR is a promising field that has been growing steadily since the 1990s. Since then, new approaches have been created, particularly in terms of the information available on the web. These SRIs are normally based on textual or multimedia content or the use of external resources (semantics). They allow the query to be matched to documents without taking the structural aspect into account. This, of course, excludes the option of including the relationships between the entities in a document. The structural aspect is primarily based on a tree-like document representation. It allows the documents to be structured in such a way so that the various relations of a document representation are highlighted.

In the following subsection, we survey a literature review about different publications, which address IR in XML documents.

2.2 Information Retrieval in XML documents

The goal of an IRS is no longer to deliver an exact response to the user's request, but rather to provide the user with a set of results arranged by relevance. Querying a collection of XML documents means comparing the query with all the XML documents in the document database. Indeed, the XML document is a semi-structured document by essence. Therefore, one of the major problems of IR is to be able to compare two documents by taking into account the content, the structure, and the semantic aspect. In our previous paper, we discussed two categories of document representation: dependent and data-independent structures.

In what follows, we present the body of work that has addressed the problem of IR in XML documents. The work [3] provides an information retrieval paradigm based on a similarity metric to rigorously compare XML documents. It sets a comparison framework of XML document structure based on the commonality of subtrees and the semantics of labels. [4] Performs IR in XML documents using the notion of tree tuples to semantically identify consistent substructures. The work [5] addresses the IR problem at the structural level by path matching between the XML document and the query representing the user need. In terms of content similarity, a matrix is generated utilizing artificial language processing techniques to compute the similarity between the keywords. For the calculation of semantic similarity, the approach relies on fuzzy matching.

In [6], the proposed approach is based on edit distance, which takes into account the content's structural and textual similarities. The suggested structural similarity algorithm merges the set of DTDs (Document Type Definition). This latter comprises the document collection into an undirected graph by employing the edit distance method and the shortest path method [7]. The structural similarity is computed preterially between the subtree of relevant nodes S and the query tree. The extraction of the S subtree starts with a selection phase of the relevant leaves. Paths are extracted based on the existing nodes in the query, from the root to the leaf. These relevant node paths are merged into a subtree. These steps reduce the size of the subtrees and increase the efficiency of the proposed model because the time of the edit distance path strongly depends on the cardinality of the input trees. In [8], the approach used for IR is based on two scores. The first one (content score) is propagated at the tree level in order to obtain the sub-trees containing the relevant leaves; the score is computed by a weighting algorithm of the form tf-idf. The second (structure score) computes the score of the subtrees previously extracted by the tree editing distance algorithm.

In the next part, we offer a state-of-the-art review of publications that have employed graphs for IR objectives.

III. GRAPH-BASED INFORMATION RETRIEVAL

3.1 Preamble

The classical graph theory problem can be described as follows: Given a graph database $D = \{G_1, \ldots, G_n\}$ and a graph representing the query q . Finding all the graphs in

which q is a subgraph, is equivalent to finding the suitable match between the query q and the D [1].

Graphs have been widely instrumental in the case of complex document representation and used in a variety of fields owing to the graph's vital function in increasing the meaning of the document represented.

Our previous research [9] has concentrated on aspects that contribute semantics to graph-based IR methods in the context of images.

Using graph theory to solve IR issues entails taking structural, contextual, and semantic factors into consideration. The combination of these aspects increases the accuracy and best meets the need described in the query. Furthermore, the graph's flexibility allows for the modeling of multiple relationships between the same nodes.

The study [10] presents a method applied at the indexing phase, in order to extract sub-graphs for IR purposes. The adapted approach consists of matching the size of the query to the size of the sub-graphs utilized to create the index. [11] uses a graph-based approach for enhanced bibliographic retrieval to a co-citation network incorporating citation context information; the method is based on a graph similarity calculation algorithm and the Random Walk with Restart (RWR) algorithm. The authors of [12] describe and structure the events of a document in order to build the text summary. The method consists in building the event graph by combining machine learning and rule-based methods. This extractive multi-document summarization approach chooses sentences based on the significance and temporal structure of events.

Graphs are made up of vertices and edges. Edges are responsible for combining and linking the vertices. The relations in the graph express the relation of membership or typing [1]. In [13], the proposed model is based on the graph which is an algebraic model closely related to the vector space model. Each vector coordinate is a value that expresses the significance of the term in the document or query. A bipartite graph is used in [14] to represent the documents with the indexed terms in the document collection. The link reflects the relationship between the document and its own indexed terms. Work [15] proposes a technique for bibliographic retrieval by an interface of interrogation for documentary bases by natural language, the structuring of the request and the documents is based on graphs.

[15] Presents a technique for bibliographic retrieval which uses an interface of interrogation for documentary bases and relies on natural language. On that regard, the structuring of the request and the documents is based on graphs.

In the next section, we discuss within the context of IR the relationships between the components of a graph.

3.2 Relationships between the components of the graph

A graph G is a set of nodes connected by links called edges. Indeed, an edge can carry information about the direction of navigation from one node to another, typing information or a content serving the user need. Moreover, the information carried by the link differs according to the context and the objective of the study. In the context of social networks, edges can express connections, friendship links between individuals. Edges on the Internet might indicate wire or wireless connections between computers or routers. As for the web, edges can reflect the hyperlinks between web pages. In rail networks, edges can be used to express connections between stations. As for road networks, edges can represent the road segments between its intersections. At documents, edges convey the relationship between the document and its own indexed terms or sometimes they express the link between nodes that correspond to the distance between documents, etc.

Following that, we divide a graph's relations into three categories: grammatical relations, string relations, and numerical relations.

- Grammatical relations

The vertices of a graph representing a document depend on the words or morphemes of a text. Syntactic functions produced by a dependent grammar are often used to designate such edges. In the literature, grammatical relations are used to structure a textual document in order to build a hierarchical structure that can be browsed and analyzed. In [16], the text is structured using a dependency parsing process. Grammatical relations, according to the same paper, are intended for sentence identification and syntactic structure creation. They play an essential role in the semantic analysis phase [17]. Independent grammar is a process that determines the type of dependent relationship that exists between the terms in the document. The work [15] proposes a method of bibliographic retrieval which uses a query interface for documentary databases and relies on natural language. In the same vein, a parser is used to structure the query and the documents, allowing it to display the many grammatical relations that link the text fragments. It is worth mentioning that this type of relationship can only be applied to IR in the text.

- String relations

The String relations include typing or membership information and describes the link's characteristics.

Moreover, this type of relation enriches the semantics of the document. Because of its great expressiveness, this type of relationship is often utilized in image and video IR. In the same context, this type of connection can express spatial and temporal relations. Allen relations [18] allow to structure the content of audiovisual sequences according to temporal information. Indeed, [18] identifies a complete set of temporal relations which can exist between two intervals. In [12], document is represented as an event graph where a graph representation involves not only the recording of events, but also the representation of temporal connections.

The work [19] illustrates the spatial interactions that govern the relationships between the image's parts. The scenes are represented by relational graphs that include information on area types and spatial layouts. The study [1] focuses on a case-based reasoning application in the field of CAD (Computer Aided Design), where cases are design items represented by directed label graphs. The work [20], [21] and [22] use spatial and temporal relations in the field of image IR.

- Numerical relations

Numerical relations are used in all fields of IR. Their basic idea is to provide a numerical value to each edge in the graph, called edge weight. The weight of an edge is computed either by a weighting function, or fixed according to the needs of the approach. The weighting function is a mathematical expression that is used to calculate sums, integrals, or averages in which certain components are more important or influential on the same set than others. In [23], the weighting function is used in the context of multi-structured documents for document classification purposes. The proposed weighting function expresses the constraints related to hierarchical or contextual. In other words, it expresses the distribution of these components in the graph and the nature of the relations between these components. In [8], the weight of an edge, abbreviated as structure score, is obtained by combining all the scores of the editing distance of the subtrees. In [5], the weight of an edge positioning in a *i* hierarchy is defined as $\frac{1}{2^i}$. This work is applied to an XML document hierarchy modeling for IR purposes. In [24], the nodes of the tree correspond to the XML documents and the relationship between the nodes correspond to the distance between the documents. Therefore, the relationship between the nodes of the structure ensures the preservation of the order. In the same regard, the tf-idf paradigm is also used to weight the relationships of the graph.

The method followed in [25] is based on a summary of the tree, in which a collection of vectors is retrieved by sequentially reducing the structure and aggregating the leaves containing the text. The weighting function is used in the same way as tf-idf-edf [26] to assign a weight to each node and edge that reflects its relevance in the collection to which it belongs. [10] presents a model for presenting terms as nodes and the number of occurrences of terms as the relationship between nodes.

IV. DISCUSSION

In this paper, we have reviewed some work that have engaged with IR in XML documents. We also discussed several studies in which graphs have been utilized to represent texts in the setting of IR. As we proceeded, we conducted research on the many sorts of connections that occur between entities. Indeed, the graph is made up of nodes and the connections between them. The relations constitute the backbone of the graph since they make explicit the nature of the link and add contextual, structural and semantic information.

Table 1 summarizes the advantages and limitations of each type of relationship.

Table 1: Advantages and limitations of each type of relationship

RELATIONS TYPE	ADVANTAGES	LIMITS
String Relations	• Well expression of characteristics • Expression of temporal and spatial relations • Expression of semantic relations • Applicable in all areas of IR	• Order not taken into account • High complexity
Grammatical relationships	• The parser is well defined • Appropriate complexity • Multiple link types generation • Semantic expression support	• Order not taken into account • Poor information is carried by the relationship • Not applicable in image and video

		retrieval
Numeric relationships	• Numeric relationships • Appropriate complexity • Applicable in all areas of information retrieval	• A weighting function must be established • Necessary interpretation of results

There are three types of graph relations: string relations, grammatical relations, and numerical relations. String relations allow users to describe the properties of the connection to express a type or membership link, as well as depict the temporal and geographical relationships between graph nodes. In contrast to grammatical relations, which are exclusively used in textual corpus to approve the nature of the grammatical relationship between the words of a textual document, they are relevant in all IR settings. Grammatical and String relations do not convey the order of the nodes in the document structure, but numerical relations do. The connection weight can represent the degree of distribution of the graph's components or the significance of a node in respect to a node in the query. In all IR situations, the weighting function may be utilized to add contextualized information and quantify the information supplied by the connection.

Finally, each type of relation has its specificity, depending on the objective and the context of where it is applied.

V. CONCLUSION

This study is a continuation of our previous work on information retrieval in semi-structured documents. We referred to a number of studies that employed graphs to represent documents in the context of XML information retrieval. In this context, we have concentrated on the relationships between graph components. Indeed, in the field of graph comparison, and notably in document IR, these connections transmit a substantial amount of information.

REFERENCES

[1] S. Sorlin, P.-A. Champin, and C. Solnon, "Mesurer la similarité de graphes étiquetés," *9èmes Journées Natl. sur la résolution Prat. problèmes NP-Complets (JNPC 2003)*, pp. 325–339, 2003.

[2] G. Salton and M. J. Mcgill, "Introduction to modern information retrieval (pp. paginas 400)." 1986.

[3] J. Tekli and R. Chbeir, "A novel XML document structure comparison framework based-on sub-tree commonalities and label semantics," *J. Web Semant.*, vol. 11, pp. 14–40, 2012, doi: 10.1016/j.websem.2011.10.002.

[4] A. Tagarelli and S. Greco, "Semantic clustering of XML documents," *ACM Trans. Inf. Syst.*, vol. 28, no. 1, pp. 1–56, 2010.

[5] Y. Dai and X. Ren, "A Hybrid Method to Evaluate Similarity of XML Document," no. January 2016, 2016, doi: 10.2991/emcs-16.2016.165.

[6] M. A. Tahraoui, K. Pinel-Sauvagnat, C. Laitang, M. Boughanem, H. Kheddouci, and L. Ning, "A survey on tree matching and XML retrieval," *Comput. Sci. Rev.*, vol. 8, pp. 1–23, 2013, doi: 10.1016/j.cosrev.2013.02.001.

[7] R. W. Floyd, "Algorithm 97: shortest path," *Commun. ACM*, vol. 5, no. 6, p. 345, 1962.

[8] C. Laitang, M. Boughanem, and K. Pinel-Sauvagnat, "XML information retrieval through tree edit distance and structural summaries," in *Asia Information Retrieval Symposium*, 2011, pp. 73–83.

[9] I. Belahyane, M. Mammass, H. Abioui, and A. Idarrou, "Graph-Based Image Retrieval: State of the Art," in *International Conference on Image and Signal Processing*, 2020, pp. 299–307.

[10] S. H. Farhi and D. Boughaci, "Graph based model for information retrieval using a stochastic local search," *Pattern Recognit. Lett.*, vol. 105, pp. 234–239, 2018.

[11] M. Eto, "Extended co-citation search: Graph-based document retrieval on a co-citation network containing citation context information," *Inf. Process. Manag.*, vol. 56, no. 6, p. 102046, 2019.

[12] G. Glavaš and J. Šnajder, "Event graphs for information retrieval and multi-document summarization," *Expert Syst. Appl.*, vol. 41, no. 15, pp. 6904–6916, 2014.

[13] Y. Champclaux, T. Dkaki, and J. Mothe, "Enhancing high precision by combining Okapi BM25 with structural similarity in an information retrieval system," *ICEIS 2009 - 11th Int. Conf. Enterp. Inf. Syst. Proc.*, vol. ISAS, pp. 279–285, 2009, doi: 10.5220/0002017202790285.

[14] Q.-D. Truong, T. Dkaki, J. Mothe, and P.-J. Charrel, "Information retrieval model based on graph comparison," *Journées Int. d'Analyse Stat. des Données Textuelles (JADT 2008), Lyon, Fr. 12-MAR-08-14-MAR*, vol. 8, pp. 1115–1126, 2008.

[15] Y. Zhu, E. Yan, and I.-Y. Song, "A natural language interface to a graph-based bibliographic information retrieval system," *Data Knowl. Eng.*, vol. 111, pp. 73–89, 2017.

[16] Z. Zhang, L. Wang, X. Xie, and H. Pan, "A Graph Based Document Retrieval Method," *Proc. 2018 IEEE 22nd Int. Conf. Comput. Support. Coop. Work Des. CSCWD 2018*, no. 61672181, pp. 660–665, 2018, doi: 10.1109/CSCWD.2018.8465295.

[17] J. Ma, "Research on Chinese dependency parsing based on

statistical methods," *Unpubl. PhD thesis, Harbin Technol. Univ.*, 2007.

[18] J. F. Allen, "Time and time again: The many ways to represent time," *Int. J. Intell. Syst.*, vol. 6, no. 4, pp. 341–355, 1991.

[19] S. Aksoy, "Modeling of remote sensing image content using attributed relational graphs," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2006, vol. 4109 LNCS, pp. 475–483, doi: 10.1007/11815921_52.

[20] S. Berretti, A. Del Bimbo, and E. Vicario, "Modelling spatial relationships between colour clusters," *Pattern Anal. Appl.*, vol. 4, no. 2, pp. 83–92, 2001.

[21] W. W. Chu, C.-C. Hsu, A. F. Cárdenas, and R. K. Taira, "Knowledge-based image retrieval with spatial and temporal constructs," *IEEE Trans. Knowl. Data Eng.*, vol. 10, no. 6, pp. 872–888, 1998.

[22] E. G. M. Petrakis, C. Faloutsos, and K.-I. Lin, "ImageMap: An image indexing method based on spatial similarity," *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 5, pp. 979–987, 2002.

[23] A. Idarrou, "Entreposage de documents multimédias: comparaison de structures." Toulouse 1, 2013.

[24] A. M. Vercoustre, M. Fegas, Y. Lechevallier, and T. Despeyroux, "Classification de documents {XML} a partir d'une representation lineaire des arbres de ces documents.," *Actes des 6eme journees Extr. Gest. des Connaissances (EGC 2006), Rev. des Nouv. Technol. l'Information*, no. 2002, pp. 433–444, 2006.

[25] S. Chagheri, C. Roussey, S. Calabretto, and C. Dumoulin, "Classification de documents combinant la structure et le contenu," in *8ème COnférence en Recherche d'Information et Applications CORIA 2012*, 2013, p. p-261.

[26] K. Sauvagnat, "Modèle flexible pour la recherche d'information dans des corpus de documents semi-structurés." Université Paul Sabatier-Toulouse III, 2005.