Pattern Recognition 72 (2017) 266–284

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Fuzzy generalized median graphs computation: Application to content-based document retrieval

Ramzi Chaieb[a,e], Karim Kalti[b,e], Muhammad Muzzamil Luqman[c], Mickaël Coustaty[c], Jean-Marc Ogier[c], Najoua Essoukri Ben Amara[d,e,*]

[a] National Engineering School of Tunis, Tunis El Manar University, Tunis, Tunisia
[b] Faculty of Sciences of Monastir, Monastir University, Monastir, Tunisia
[c] L3i Laboratory, University of La Rochelle, La Rochelle, France
[d] National Engineering School of Sousse, Sousse University, Sousse, Tunisia
[e] LATIS – Laboratory of Advanced Technology and Intelligent Systems, ENISo, Sousse University, Sousse, Tunisia

ARTICLE INFO

Article history:
Received 28 November 2016
Revised 20 April 2017
Accepted 27 July 2017
Available online 29 July 2017

Keywords:
Fuzzy attributed relational graph
Graph embedding
Fuzzy set median graph
Fuzzy generalized median graph
Similarity measure
Document image retrieval

ABSTRACT

Fuzzy median graph is an important new concept that can represent a set of fuzzy graphs by a representative fuzzy graph prototype. However, the computation of a fuzzy median graph remains a computationally expensive task. In this paper, we propose a new approximate algorithm for the computation of the Fuzzy Generalized Median Graph (FGMG) based on Fuzzy Attributed Relational Graph (FARG) embedding in a suitable vector space in order to capture the maximum information in graphs and to improve the accuracy and speed of document image retrieval processing. In this study, we focus on the application of FGMGs to the Content-based Document Retrieval (CBDR) problem. Experiments on real and synthetic databases containing a large number of FARGs with large sizes show that a CBDR using the FGMG as a dataset representative yields better results than an exhaustive and sequential retrieval in terms of gains in accuracy and time processing.

1. Introduction

During the recent years, the widespread use of digital devices (smartphones, cameras, tablet computers, etc.) and the tremendous growth in digitizing various collections of documents (invoices, books, historical documents, etc.) have resulted in the creation of large databases of document images. Browsing and retrieving from such large-scale document images is more than ever a challenging problem in data mining research. There is still a growing need for efficient and fast methods for searching information in these huge databases. Traditional approaches for Document Image Retrieval (DIR) are text-based [1–3]. However, fully text-based approaches are not practical enough for DIR applications due to many limitations such as time consumption, subjectivity of the annotator, ambiguity (same words but not in the same place in a document image), etc. Therefore, Content-based Document Retrieval (CBDR) approaches have been proposed to overcome the shortcomings of the text-based retrieval [4–10]. Most document image content representation methods are based on the vector space model of information. One of the most important advantages of this representation model is the computation time and memory consumption. However, this popular method of document representation does not capture important structural information, such as the location of an object within the document image or the spatial relationships between document image regions. Also, vector based representations suffer from the constraint of fixed dimensionality of feature vectors [11]. Therefore, structural signature represents some interesting alternatives to construct faith documents representation, through the use of document image regions and their topological relationships. Indeed, graph-based document image representation model can preserve document image structural and spatial information and furthermore it doesn't require a predefined size of the representation. It provides a rich and holistic description of the layout and content of the analyzed document images. It was shown to outperform the traditional vector representation for several applications in many fields [11,12]. However, graph-based representations are usually computationally more expensive than vector-based representations as they require exponential time and space due to the NP completeness of the problem. For further read-

* Corresponding author.
E-mail addresses: ramzi.chaieb@hotmail.com (R. Chaieb), karim.kalti@gmail.com (K. Kalti), muhammad_muzzamil.luqman@univ-lr.fr (M.M. Luqman), mickael.coustaty@univ-lr.fr (M. Coustaty), jean-marc.ogier@univ-lr.fr (J.-M. Ogier), najoua.benamara@eniso.rnu.tn, benamaranajwa@gmail.com (N. Essoukri Ben Amara).

http://dx.doi.org/10.1016/j.patcog.2017.07.030
0031-3203/

ing on graph-based representations and applications we refer the interested reader to [11,12].

Due to many particularities of document images (for example: noise and degradation, overlapping layouts, presence of handwriting, etc.), eventual segmentation errors may occur after the segmentation of document regions. Segmentation results are highly dependent on the performance of the segmentation algorithm which might be unstable over various document images. Therefore, the use of fuzzy graph-based description (instead of a classic graph-based representation which is a rigid description) could be helpful to add flexibility against these errors. Moreover, a document image content may generally be submitted to variations that introduce some vagueness or uncertainty in the way to describe the information. Therefore, representing the content by fuzzy graphs allows to capture the maximum information from a document image with a certain error-tolerance. Structural and visual features should be represented by fuzzy concepts, such as "Near" and "Far", "Big" and "Small", etc. These concepts are described with the use of the fuzzy set theory. A crisp description can be represented as a special case of a fuzzy description. For further reading on fuzzy graphs and its applications we refer the interested reader to [13–15].

Generally, the CBDR process is composed of three main phases: extracting features, structuring feature space and retrieving. The first two phases are usually performed off-line. In this paper, we assume that all document images are represented by Fuzzy Attributed Relational Graphs (FARGs). We will focus on the second phase which aims at organizing the input fuzzy graphs into an efficient data structure in order to improve and accelerate retrieval results. The proposed approach avoids the sequential search in a FARG database by direct access to a reduced set containing the FARGs most similar to a query FARG. Therefore, we used the concept of median graph for the structuring phase of a CBDR system. Median graph aims at representing an input set of graphs with no constraint on the set size. It is frequently used to indicate the graph that best captures the information presented in all input graphs. In other words, the median graph of a given set of graphs is the graph that minimizes the sum of distances to all other graphs in this set. Median graphs have wide-spread applications in diverse fields such as pattern recognition, classification, image analysis, etc. [16–19]. In recent years, more and more research efforts have been devoted to median graph problem. A brief review of some strategies for median graph construction will be outlined in the related work section.

In this study, we propose a new algorithm for the computation of the Fuzzy Generalized Median Graph (FGMG) in order to improve the speed and accuracy of retrieval processing. The first contribution of this paper consists in representing query and database document images by FARGs. One motivation for the FARG representation is to keep advantages from both the graph and the vector domains (power of representation of graphs and easiness manipulation of the vector representations). Besides, FARGs allow to represent document images with a fuzzy approach which is similar to human perception [9,10,13]. FARGs provide both syntactic and semantic information. Syntactic information is held by the layout of the graph (nodes and edges), while semantic information is expressed by attributes associated to nodes and edges in the graph. Describing images by exploiting these two informations will reduce the semantic gap between low-level features and high-level concepts and therefore improve the retrieval results. Also, the FARG representation is very useful for reducing the effect of possible segmentation errors which may be occurred after the segmentation phase [13]. The second contribution of this paper is a new algorithm for the computation of the FGMG based on FARG embedding into a vector space. In the context of dealing with a large mass of document image datasets, it is not trivial to perform an exhaustive and sequential comparison of the query with all document images in the database due to the high computational complexity requirements. Thus, we have developed a new FGMG computation algorithm in order to contribute to the structuration of the FARG space. The third contribution of this paper is a new FARG embedding method in order to reduce the computation time of the FGMG. FGMG computation is based on computing the distance between every pair of FARGs. Since all the possible combinations of FARGs need to be explored, the computation of the FGMG will be therefore exponential in the number and size of input FARGs. Embedding FARGs into a vector space solves this problem since FARGs are represented by feature vectors [14,20,21,22]. Thanks to this new method, we are able to keep power of representation of FARGs while manipulating the vector representation of the FARGs.

The remaining of this paper is organized as follows. The next section gives some preliminaries for the new FGMG algorithm and a detailed presentation of the concept of the median graph. Section 3 presents a literature review of related work. Section 4 describes the FGMG computation algorithm in detail. Section 5 introduces an example of application of FGMGs to the CBDR problem. Experimental results are evaluated and discussed in Section 6. Finally, concluding remarks and future work are given in the last section of this paper.

2. Fuzzy median graph theory

In this section, we present some preliminaries for the new FGMG algorithm, including the definition of the Fuzzy Attributed Relational Graph (FARG) and those of a Fuzzy Set Median Graph (FSMG) and of a Fuzzy Generalized Median Graph (FGMG).

2.1. Fuzzy attributed relational graph

A FARG is a graph whose nodes (also called vertices) and edges (also called arcs) are both represented with fuzzy attributes. Given a finite fuzzy attribute set A, a FARG G can be defined as a quadruple (N, E, μ, υ) where N is a non-empty finite set of nodes, $E \subseteq N \times N$ is the set of edges, $\mu: N \rightarrow A$ is a function that associates a fuzzy attribute value in A to each node, and $\upsilon: E \rightarrow A$ is a function that associates a fuzzy attribute value in A to each edge. In this paper, FARG nodes represent document image regions and FARG edges represent spatial relationships between the regions.

2.2. Fuzzy median graph

Given a set of graphs, the median graph is frequently used to indicate the graph that best represents the set. In other words, the median graph is the graph that best captures the information presented in all graphs. Intuitively, the median graph is located in the center of the given graph set. Basically, two different definitions for median graphs have been presented: the set median graph and the generalized median graph. One difference between them is in the search space of graphs where the median is looked for. The generalized median graph is usually constructed from a larger set of graphs.

2.2.1. Fuzzy set median graph

Let U be the set of FARGs that can be constructed using a given set of attributes A. Given a set $S = \{G_1, G_2, \ldots, G_i\} \subseteq U$, the Fuzzy Set Median Graph (FSMG) of S is defined as the FARG $\hat{G} \in U$ that minimizes the Sum Of Distances (SOD) to all FARGs in S.

$$\hat{G} = \underset{G \in S}{\operatorname{argmin}} \sum_{i=1}^{|S|} d(G, G_i) = \underset{G \in S}{\operatorname{argmin}}\, SOD(G) \tag{1}$$

where $d(G, G_i)$ denotes a distance or a dissimilarity measure between a candidate median FARG G and a FARG $G_i \in S$.