Analysis and Evaluation of Current Graph-Based Text Mining Researches

Jae-Yong Chang and Il-Min Kim
Dept. of Computer Engineering, Hansung University
2-Ga SamSun-Dong, SungBuk-Gu, Seoul, 136-792, Korea
{jychang, ikim}@hansung.ac.kr

Abstract. Traditionally, since text mining has been assumed unformatted text (document), it is necessary to represent text as a simplified model. One of most commonly used model is the vector space model, in which text is represented as a bag of words. Recently, much research tried to apply a graph-based text model for representing semantic relationships between words. In this paper, we survey research trends of graph-based text representation models for text mining. We summarized each model, its features and forecasted further researches.

Keywords: Text Mining, Vector Space Model, Text Representation, Graph Model

1 Introduction

Text mining is an area of data mining and its goal is to analyze unformatted text and finding out the hidden knowledge of the text. Traditionally, one of the most commonly used models for representing text in text mining is VSM(Vector Space Model) [1], in which frequently appeared words and their weights are expressed as vector forms. In this paper, we will analyze the research trends of text mining which are based on graph. At first, we delve into researches on the vector space models, and then, we will systematically summarize the graph-based text models according to their characteristics.

2 Vector Space Model (VSM)

In VSM, a word is represented as a dimension, and a text is represented as a point in an n-dimensional space. Since VSM can represent an unformatted text document as simple and formulaic notation, various algorithms which had been used in data mining
can be applied without any modification. Because of the advantage, lots of researches on VSM are being actively carried out. However, VSM also has following disadvantages because of its simplicity [2].

- If two documents have similar meaning but they used different words, their similarity cannot be computed easily.
- The meaning of a text or the structure of a text cannot be expressed in VSM.
- Since each word is independent from others, word appearance sequence or other relations cannot be represented in VSM.

Various researches have been carried out in order to solve these problems. So far, graph-based text representation model is known as one of best solution for these problems.

3 Classification of Graph-based Model

3.1 Classification by Node Representation

A graph $G$ is derived from a text or a text set can be expressed as follows:

$$G = \{ V, E \}$$

In this notation, $V$ represents the set of nodes, $E$ represents the set of edges. With this notation, we can represent various graphs models according to the definitions of $V$ and $E$. In this paper, we classified the graphs in detail according to the node representation and the edge representation.

3.1.1 Node Representation Method

Nodes of graph $G$ represent text components, for examples, words, sentences, paragraphs, and texts itself. Also, nodes can represent concepts which can be considered as semantic components. According to the definition of the model, a node can represent one component, or more than two components. If a node represents one component, it is called homogenous representation. If a node represents more than two components, it is called heterogeneous representation. In addition, if nodes can be either weighted or unweighted according to whether or not to assigning weighted value into the nodes.

- **Homogenous representation vs. heterogeneous representation**

  Commonly used notations used in homogenous representations represent words as nodes [3-8]. For many cases, co-occurrence information between words is expressed in graphs. Co-occurrence means that two words are appeared more than once within a
sentence or n-window size in a document, in that case you need to connect the words with edge. Grammatical associations between words or semantic similarities also could be represented as a homogeneous graph model [10]. Since this representation is simple, the cost for building the model and analysis is low. Another advantage of this representation is that the existing algorithms used in the vector space model can be applied without modification. In some researches, they also used homogenous representations, in which sentences, paragraphs or concepts are represented as nodes [12].

► **Weighted and Unweighted**

In weighted representations, weighted value has been assigned in each node. In contrast, a node does not have weighted value in unweighted representations. Most researches assume weighted nodes which indicate the importance of the node in the graph. In order to evaluate the weight of nodes, some researches, for example PageRank [19], exploited the number of edges, the weights of edges, or the weight of nodes which are connected by the edge.

### 3.1.2 Edge Representation Method

An edge represents the relationship between nodes. Edges can be classified into three categories according to their appearances. Each category will be explained in detail.

► **Directed vs. Undirected edges**

Directed edges are used for indicating node orders or mutual interactions. For example, if you want to indicate the order of the words, you need to use directed edges in the graph [4]. Each word in a sentence has its grammatical role (for example, subject, verb, and object), you need to use directed edges [9, 10]. If you use a tree, it could be considered a directed edge representation [13].

If there are no orders or mutual interaction between nodes, undirected edge is used for connecting related nodes. When you need to represent the co-occurrences between words, you use the undirected edges in the graph [6, 18].

► **Weighted vs. Unweighted edges**

Weights in edges score numbers which indicate the degree of node relationships. For example, the weight could be the number of co-occurrences which relate words appear at the same time in a co-occurrence graph [3, 4]. Weight value can be the distance of two words in a text. If two nodes are not quantitatively related, unweighted edges are used [5, 8, 9, 10].

► **Labeled vs. Unlabeled edges**

Labeled edges were used in some graph models [5, 7, 9, 10]. Labels represent the roles of the nodes, for many cases, they indicate the relationships between words. In the paper [10], the edge is labeled 'verb' when the edge connects from the 'subject' node to the 'object' node. In the paper [9], a sentence is represented as a parsing tree. The label depicts the part of speech of each word. When each word's grammatical role needs to be indicated, labeled edges are generally used. The other cases, unlabeled edges are commonly used.
4 Conclusions and Further Research Trends

This paper explained and classified the previous text-representing models based on graphs. As we surveyed in this paper, graph-based text representing models adopted different modeling according to their goals and application areas. In other words, there were no efforts to standardize the graph models. Therefore more systematic research on graph model for representing text is required. The systematic research for standard graph model will be useful for document classification, aggregation, summary, search, and will be applied for various document analyses.

References