The Composite Graph Model for Web Document and its Impacts on Graph Distance Measurement

Kaushik K. Phukon

Department of Computer Science, Gauhati University
Guwahati-14, Assam, India
kaushikphukon@gmail.com

Abstract

It has been accepted that the composite graph model can represent a web document with minimum loss of information .The composite graph model [8] is the combination of the TSGM and CSGM. With the help of the composite model it is possible to hold more information about a web page than ever. In this article we are going to put forward some new aspects of the composite model and graph distance measure for web documents. The methods presented in this paper are not information retrieval systems .We do not select web documents for retrieval; we are trying to measure the dissimilarity between/among them for the purpose of clustering or classifying them through the application of machine learning techniques.

Keywords*: Graph, information, web document, distance*

1. Introduction

Targeting useful and relevant information on the WWW is a topical and highly complicated research area. A thriving research effort that feeds into this area is document clustering, which closely related with areas usually known as text classification and/or text categorization. Today, the WWW represents one of the largest, distributed, heterogeneous, semi-structured repositories of multimedia content. The state-of-the-practice today is to use existing search engines to provide search functionality to the user. However, typical queries elicit hundreds, sometimes even thousands, of URLs from search engines, forcing the user to wade through them in order to find the URL(s) she needs. In large part, this limitation of search technology can be attributed to the following [10]:

Polysemy: the words involved in the search have multiple meanings. For example, a user searching for windows may be interested in either the operating system or the physical artifact.

Phrases: a phrase may be different from words in it. e.g., the meaning of the phrase "partition magic" (a disk partition management tool) is quite different from the meaning of the individual words "partition" and "magic".

Term dependency: words in the terms are not totally independent of each other. For example, a user may look for details about a product made by a particular company and type in Sun's Enterprise Computer Series. Obviously, each word in this term is dependent on each other.

These problems are independent of how good the algorithms that associate keywords with the contents of a page are. One possible solution to this problem is the use of graph theoretic ideas.

Conventional document representation methods consider documents as vase of words and ignore the meanings and ideas their authors want to convey. It does not capture important structural information, such as the order and proximity of word occurrence or the location of a word within the document. It also makes no use of the mark-up information that can be easily extracted from the web document HTML tags. It is this deficiency that causes similarity measures to fail to perceive contextual similarity of web documents [7].

A graph G is a 4-tuple: $G= (V, E,\alpha,\beta)$, where V is a set of nodes (vertices), $E \subseteq V \times V$ is a set of edges connecting the nodes, $\alpha : V \rightarrow \Sigma v$ is a function labeling the nodes, and $\beta : V \times V \rightarrow \Sigma e$ is a function labeling the edges (Σv and Σe being the sets of labels that can appear on the nodes and edges, respectively). For brevity, we may refer to G as $G= (V, E)$ by omitting the labeling functions.

Web document clustering methodologies can generally be classified into one of three distinct categories [1,3].In the clustering based on web content we study the actual content of web pages and then apply some method to learn about the pages. In general this is done to organize a group of documents into related categories. This is especially beneficial for web search engines, since it allows users to more quickly find the information they are looking for in comparison to the usual infinite ordered list. In the clustering based on web usage the goal is to examine web page usage patterns in order to learn about a web system's users or the relationships between the documents on the basis of association rules created from web access logs, which store the identity of pages accessed by users along with other information such as when the pages were accessed and by whom. In the third category of web clustering methodologies, clustering based on structure, we examine only the relationships between web documents by utilizing the information conveyed by each document's hyperlinks.

In this paper we are concerned only with the clustering based on web content.

Several methods are there for representing web document content (or text documents in general) as graphs. But none of them are well established as a de facto standard for representing web documents as graphs. We are trying to establish the composite model[8] as the standard one.

2. The Tag Sensitive Graph Model

Under the Tag Sensitive Graph representation each unique term appearing in the document becomes a node in the graph representing that document [1,4,5,8]. Each node is labeled with the term it represents. The node labels in a document graph are unique, since a single node is created for each keyword even if a term appears more than once in the text. Second, if word *a* immediately precedes word *b* somewhere in a "section" *s* of the document, then there is a directed edge from the node corresponding to term *a* to the node corresponding to term *b* with an edge label *s*. An edge is not created between two words if they are separated by some sentence terminating punctuation marks (periods, question marks, and exclamation points).

Sections we have defined for HTML documents are: *head,* which contains the title of the document and any provided keywords; *link,* which is text that appears as hyper-links on any section of the web document; *address* which also contains valuable information; and *text,* which comprises the readable text in the web document (this includes text inside the body section excluding link text and address). The edges are labeled according to head (H), link (L), address (A) or text (T). We always create an edge between first elements of the head section and the first element of the address section and label this edge as ‘A’.

Figure 1 is the graph representation of a web page which is taken from the University of South Florida web site (http://www.usfdiningservices.com/mealplans.php - document 1) created using the TSGM (i.e. standard) model [1] (The page may not be available currently).

This method emphasizes on representing web documents on the basis of the sections and is capable of utilizing the markup information available in the web document. It can capture some important structural information such as the location of a word within a document. Being a directed graph it can represent the sequence of word occurrence within a document.

This model cannot reflect the proximity of words directly. Further calculations have to be made to know the distance between word pairs. This leads to reduced accuracy to perceive contextual similarity of web documents due to the variation of words the documents contain.

Also there will be a large number small of disconnected components. These small components will be totally ignored by the MCS computing algorithm irrespective of the type and kind of the algorithm, resulting the loss of important information. To retain all the information in the process of computing the MCS, the graph must be a connected one which is not possible in this model of graph representation.

3. The Context Sensitive Graph Model

Under the Context Sensitive Graph representation also each unique term appearing in the document becomes a node in the graph representing that document; but there is a user-provided parameter, '*n*'. Instead of considering only terms immediately following a given term in a web document, we look up to *n* terms ahead and connect the succeeding terms with an edge that is labeled with the distance between them (unless the words are separated by certain punctuation marks or they are in a different section of the web page)[1,8].

Being a directed distance graph, it can retain information about word pairs which are at a distance of at most 'n' in the underlying document where 'n' is the order of the graph. It can hold almost all the information that we require to analyze or cluster ordinary documents.

This method is not suitable for web documents because a web document is much different than a general document as it contains various markup information. This method also fails to create an almost connected graph for an entire web page.

4. The Composite Model for Representing Web Documents

In this representation we are using the TSGM model to represent three sections namely head, link and address because these three sections are comparatively much smaller than the text section and TSGM is capable of representing small section more efficiently than that of CSGM. Use of TSGM will enable us to utilize the markup information available which will not be possible if we use CSGM.

We are using CSGM to represent the text section because of its efficiency to represent large text section. If we use TSGM to represent this section also then there will be a loss of information, which otherwise can be used to measure contextual similarity. For the text section, the information about the proximity of words is more important than that of the markup information.

The graph in Figure1 was created by applying the TSGM model and it can be seen that several small disconnected components are there. These small disconnected components will result in reduced efficiency in determining graph distance. To have almost a connected graph for an entire web page we are proposing the following modification to our previous composite model [8].

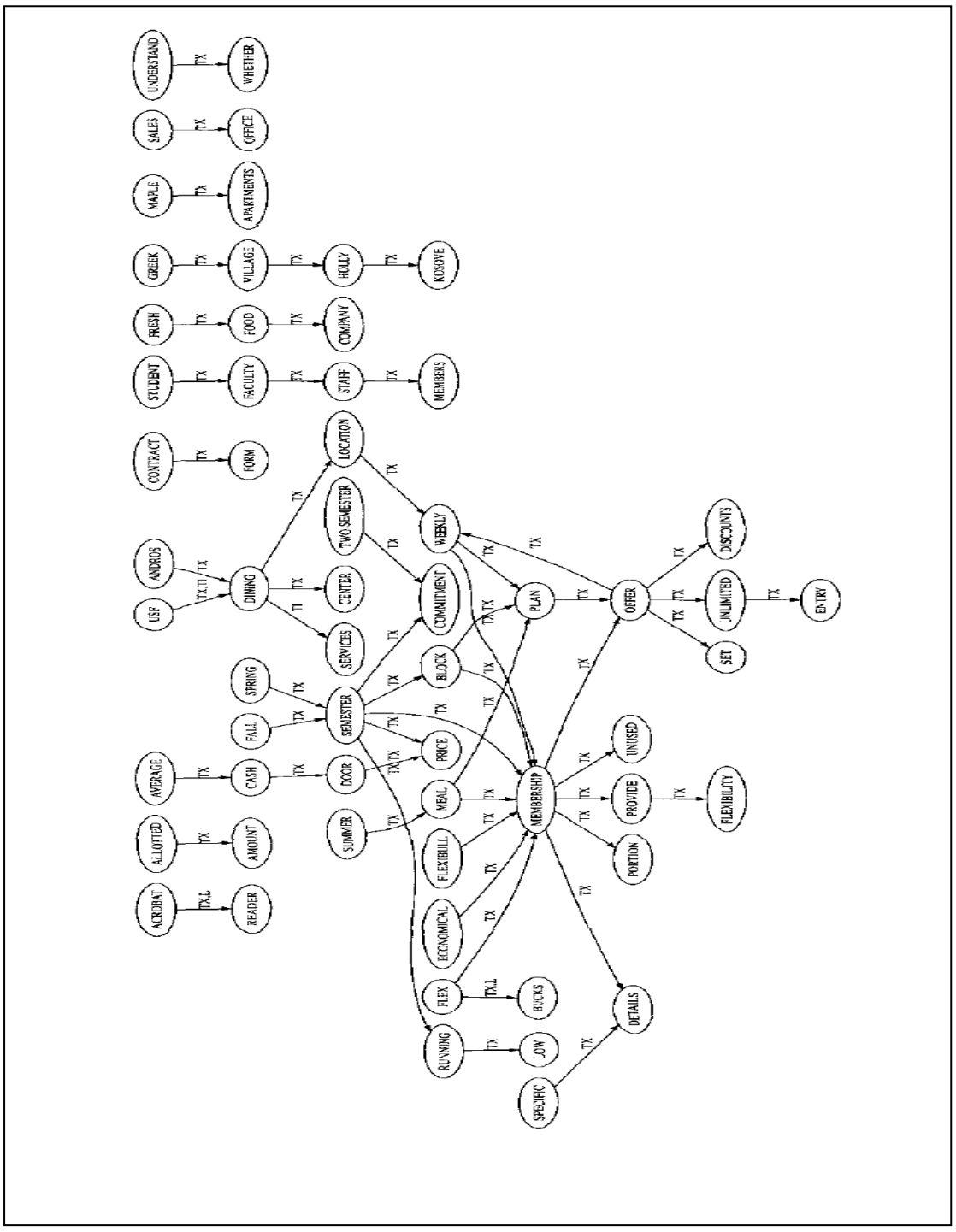

Figure 1. A Graph Created from the Document 2 using the TSGM (Standard) Representation with Isolated Nodes Omitted

If the graph created by applying the composite model contains several small disconnected components due to the sentence terminating punctuation marks; then to have the full

information utilization we may create a second order edge from the word just preceding the last word of any sentence to the first word of the immediate next sentence.

4. Impact of the Composite Model on Distance Measures Based on MCS

To have full benefit from the composite model we have modified the general graph distance measure based on MCS (Maximum Common Subgraph) as blew [9]-

$$dist_{MCS}(G_1,G_2)=\sum d^{\pm}(\mathrm{mcs}(G_1,G_2)) \;/\; \max(\sum d^{\pm}(G_1),(\sum d^{\pm}(G_2))$$

where G_1 and G_2 are graphs, $mcs(G_1,G_2)$ is their maximum common subgraph, max(...) is the standard numerical maximum operation, $\sum d^{\pm}$ is the sum of in-degree and out-degree of the vertices of the $mcs(G_1,G_2)$ [8]. This modification has been made to make use of all the information that we are capturing with the help of the composite model.

The older MCS technique for determining graph distance is unable to use all the information that the composite model can hold and hence a complete new approach for measuring graph distance have been developed. The following is the most prominent reason for developing a new technique.

The prevalent MCS technique for determining graph distance is as follows:

$$dist_{MCS}(G_1,G_2)=1-\frac{|mcs(G_1,G_2)|}{\max(|G_1|,|G_2|)}$$

where G_1 and G_2 are graphs, $mcs(G_1,G_2)$ is their maximum common subgraph, max(...) is the standard numerical maximum operation, and |...| denotes the size of the graph.

This method is considering only the number of nodes forming the MCS for calculating the distance between the graphs. For the TSGM (i.e. standard) model this method of determining the distance is very suitable because the nodes cannot hold more information that can be used by the method. In case of the composite model each node is capable of holding much more information than the TSGM model. If we use the prevalent MCS technique for determining graph distance then we hardly have any benefit from the composite model. In the newly developed technique, instead of considering only the number of nodes forming the MCS we are considering the degree of each node for calculating the distance. The importance of each node will be best explained by its degree and since we are considering directed graph, it will be very fruitful. The efficiency of this method has already been explained with the help of an example in [8].

Some other methods are also there based on MCS (Maximum Common Subgraph/Minimum Common Subgraph) as stated below-

A distance measure which has been proposed by Jonathan D. Wallis based on the idea of graph union [1], is:

$$d_{WGU}(G_1,G_2)=1-\frac{|mcs(G_1,G_2)|}{|G_1|+|G_2|-|mcs(G_1,G_2)|}.$$

By "graph union" we mean that the denominator represents the size of the union of the two graphs in the set theoretic sense; specifically adding the size of each graph ($|G_1| + |G_2|$) then subtracting the size of their intersection ($|\mathrm{mcs}(G_1,G_2)|$) leads to the size of the union (the reader may easily verify this using a Venn diagram). This distance measure behaves similarly

to MCS. The motivation for using graph union in the denominator is to allow for changes in the smaller graph to exert some influence over the distance measure, which does not happen with MCS. This measure was also demonstrated to be a metric, and creates distance values in [0,1].

A similar distance measure proposed by Prof. Em. Dr. Horst Bunke which is not normalized to the interval [0, 1] is:

$$d_{UGU}(G_1, G_2) = |G_1| + |G_2| - 2 \cdot |mcs(G_1, G_2)|.$$

Fernandez and Valiente have proposed a distance measure based on both the maximum common subgraph and the minimum common supergraph:

$$d_{MMCS}(G_1, G_2) = |MCS(G_1, G_2)| - |mcs(G_1, G_2)|,$$

where $MCS(G_1,G_2)$ is the minimum common supergraph of graphs G_1 and G_2. The concept that drives this distance measure is that the maximum common subgraph provides a "lower bound" on the similarity of two graphs, while the minimum common supergraph is an "upper bound". If two graphs are identical, then both their maximum common subgraph and minimum common supergraph are the same as the original graphs and $|G_1| = |G_2| = |MCS(G_1,G_2)| = |\text{mcs}(G_1,G_2)|$, which leads to $d_{MMCS}(G_1,G_2) = 0$. As the graphs become more dissimilar, the size of the maximum common subgraph decreases, while the size of the minimum common supergraph increases. This in turn leads to increasing values of $d_{MMCS}(G_1,G_2)$.For two graphs with no maximum common subgraph, the distance will become $|MCS(G_1,G_2)| = (|G_1| + |G_2|)$. MMCS has also been shown to be a metric, but it does not produce values normalized to the interval [0,1], unlike the MCS or WGU. Note that if it holds that $|MCS(G_1,G_2)| = |G_1| + |G_2| - |mcs(G_1,G_2)| \; \forall \; G_1,G_2$, we can compute $d_{MMCS}(G_1,G_2)$= as $|G1| + |G2| - 2|\text{mcs}(G_1,G_2)|$.This is much less computationally intensive than computing the minimum common supergraph. We can also create a version of this distance measure which is normalized to [0,1] as follows:

$$d_{MMCSN}(G_1, G_2) = 1 - \frac{|mcs(G_1, G_2)|}{\max(|MCS(G_1, G_2)|)}.$$

All the above distance measures are based on MCS and to make use of the information held by our composite model, all these measures may be modified to incorporate the vertex degree instead of simply taking the number of nodes in the MCS.

5. Conclusion

The composite method of web document representation takes into account additional web-related content information which is not done in traditional information retrieval models. It can hold almost all the necessary information such as the order, proximity of word occurrence, markup information and location of a word within a document. This model along with the enhanced distance measure is giving an increased effectiveness in the graph distance measure even though the MCS is same in both the cases [8].

Here in this paper we are suggesting one more enhancement to the composite model which may solve the issue of small disconnected components and will definitely increase the efficiency of the graph distance measure in general. The impact of the composite model on

graph distance measurement is also very impressive. We suggest all who are working in this field to adopt the composite model and the changes required in the formulae which is being used to calculate the graph distance, to have the benefits of the composite model.

References

[1] A. Schenker, H. Bunke, M. Last and A. Kandel, "Graph Theoretic Techniques for Web Content Mining", Series in Machine Perception and Artificial Intelligence — vol. 62 Copyright © 2005 by World Scientific Publishing Co. Pte. Ltd., **(2005)**.

[2] D. Lopresti and G. Wilfong, "Applications of graph probing to web document analysis", Proceedings of the 1st International Workshop on Web Document Analysis (WDA2001), pp. 51–54, **(2001)**.

[3] X. He, H. Zha, C. H. Q. Ding, H. D. Simon, "Web document clustering using hyperlink structures", Computational Statistics & Data Analysis, pp. 19-45, **(2002)**.

[4] A. Schenker, M. Last, H. Bunke and A. Kandel, "Classification Of Web Documents Using Graph Matching", presented at IJPRAI, pp.475-496, **(2004)**.

[5] A. Markov, M. Last and A. Kandel, "Fast Categorization of Web Documents Represented by Graphs", in Proc. WEBKDD, pp. 56-71, **(2006)**.

[6] J. G. Augustson and J. Minker, "An Analysis of Some Graph Theoretical Cluster Techniques", presented at J. ACM, pp. 571-588, **(1970)**.

[7] K. Shaban, "A Semantic Graph Model for Text Representation and Matching in Document Mining", PhD thesis, Electrical and Computer Engineering, Faculty of Engineering, University of Waterloo, Canada, **(2006)**.

[8] N. Deo, "GRAPH THEORY with Applications to Engineering and Computer Science", PHI Learning Private Limited.ISBN-978-81-203-0145-0.

[9] H. Bunkea and K. Shearer, "A graph distance metric based on the maximal common subgraph", Pattern Recognition Letters, vol. 19, pp. 255–259, **(1998)**.

[10] A. Joshi, Z. Jiang, "Retriever: Improving Web Search Engine Results Using Clustering ".