A Composite Graph Model for Web Document and the MCS Technique

Kaushik K. Phukon

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

Abstract

It has been accepted that a graph can represent any document with minimum loss of information. In this article we are going to put forward some new standards of graph representation and graph distance measure for web documents. With the proposed enhanced method of graph representation and distance measure we would be able to hold more information than usual and hence classify them more efficiently.

Keywords*: Graph, information, web document, tag, context, distance, subgraph*

1. Introduction

With the exponential growth of the amount of content on the Internet, the need to manage these documents also grows. Information overload is the result of this explosive growth. The volume of information available only through the Internet, presents a non-trivial real problem. This information overload can lead to psychological, physical and social problems, especially to the knowledge workers whose jobs mainly involve dealing with and processing information. In a world-wide survey conducted by Reuters News Agency, it was found that two thirds of managers suffered from increased tension and one third from ill-health because of information overload [7, 8]. It was also concluded that other effects of too much information can cause anxiety, poor decision-making, difficulties in memorizing and remembering, and reduced attention span.

Clustering and classification have been useful and active areas research that promises to help us cope with the problem of information overload on the Internet. With clustering the goal is to separate a given group of data items into groups called clusters such that items in the same cluster are similar to each other and dissimilar to the items in other clusters.

Web document clustering methodologies can generally be classified into one of three distinct categories [1]:

a) Based on Content

b) Based on Usage

c) Based on Structure

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. Web usage mining is useful for providing personalized web services, an area of web mining research that has lately become active. It helps to tailor web services, such as web search engines, to the preferences of each individual user.

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. Like the clustering based on web usage stated above, the other content of the web pages is often ignored. A graph model may be utilized to represent web page structure, where nodes in the graphs are web pages and edges indicate hyperlinks between pages. By examining these graphs it is possible to find documents or areas of interest through the use of certain graph-theoretical measures or procedures.

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

2. Representation of a Web Document using Graph

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.

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. In the present paper we are considering two experimentally established fundamental models which we will use to develop a composite model.

2.1. The Two Fundamental Models

2.1.1. Tag Sensitive Graph Model (TSGM): This model is in accordance to the standard model of document representation [1,4,5] The two changes we are proposing for this method are- one additional section *address(A)* which also contains valuable information and the nomenclature as described in the section 2.2, Figure 1.

2.1.2. Context Sensitive Graph Model (CSGM): This model is in accordance to the n-distance model of document representation [1]. The only change we are proposing for this model is its nomenclature as described in the section 2.3, Figure 2.

Both of these methods are based on examining the terms on each web page and their adjacency. Terms can be extracted by looking for runs of alphabetical characters separated by spaces or other types of common punctuation marks. Once the terms are extracted, we can use

several steps to reduce the number of terms associated with each page to some representative set.

Here, in this paper, we do not concern about information retrieval systems. Our aim is to classify web documents with the help of graphs and empower it as a primary way to represent web documents.

2.2. 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. 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 certain punctuation marks (such as periods).

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'.

An example of this type of graph representation is given in Fig-1. The document represented by the example has the title "Gauhati University", a link whose text reads "Other Universities In Assam", an address that contain "Powered by xyz" and text containing "Gauhati University Secures 26th In All India Ranking".

2.2.1. Merits and Demerits of TSGM : 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.

2.3. 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). An example of this type of graph representation is given in Fig-2. The document represented by the example has the title "Gauhati University", text containing "Gauhati University Secures 26th In All India Ranking", a link whose text reads "Other Universities In Assam", an address that contain "Powered by xyz".

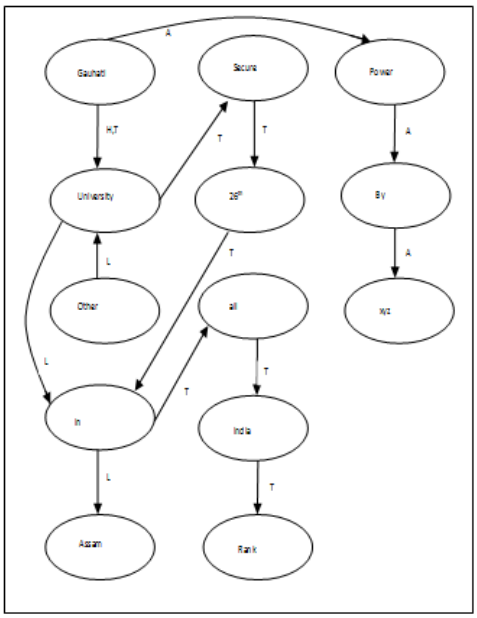

Fig 1: TSGM Representation

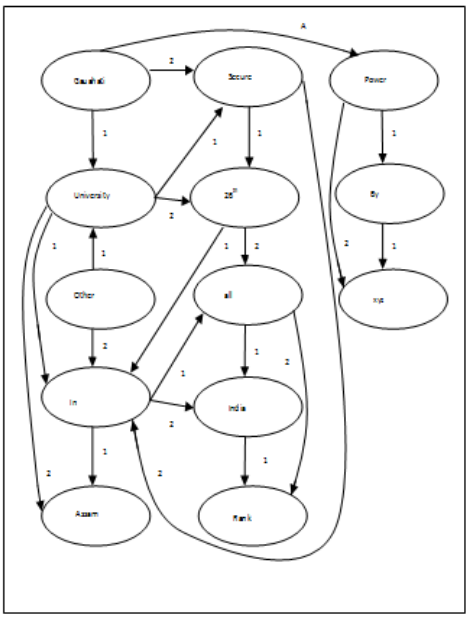

Fig 2: CSGM Representation

2.2.1. Merits and Demerits of CSGM: 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.

3. The Proposed Composite Model for Representing Web Documents

In view of the advantages and disadvantages of the two models discussed above we here propose a composite model for representing web documents on the basis of the above two fundamental approaches as illustrated below(fig.3) with the help of the same web page. i.e. The document represented by the example has the title "Gauhati University", a link whose text reads "Other Universities In Assam", an address that contain "Powered by xyz" and text containing "Gauhati University Secures 26th In All India Ranking".

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.

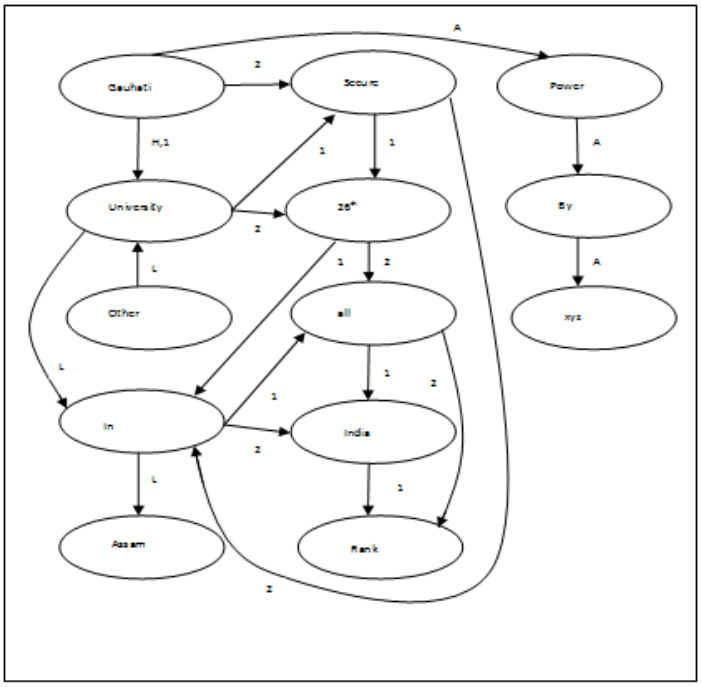

Fig 3: The Proposed Composite Representation (G_1)

This model has the advantages but does not have the disadvantages of both the parent models.

4. Graph Distance Measures

After representation stage documents (graphs) have to be compared for similarity measures. For similarity measures also there are no reported findings to indicate a de facto standard. Although the maximum common subgraph approach is a widely accepted graph distance similarity measure as stated below.

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

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. The size of a graph can be taken as the number of nodes and edges contained in the graph. In case of our proposed composite model the computation of *mcs* can be accomplished in polynomial time due to the existence of unique node labels in the considered application. In [4] it was experimentally proved that the graph method outperformed the vector method in terms of execution time in some cases.

Now to have full benefit from the proposed composite model we are enhancing the MCS distance measure as blew-

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

where $\sum d^{\pm}$ is the sum of in-degree and out-degree of the directed graph[9]. This modification has been proposed to make use of all the information that we are capturing with the help of the composite model. To see the effect of this modification let us assume a web document as below.

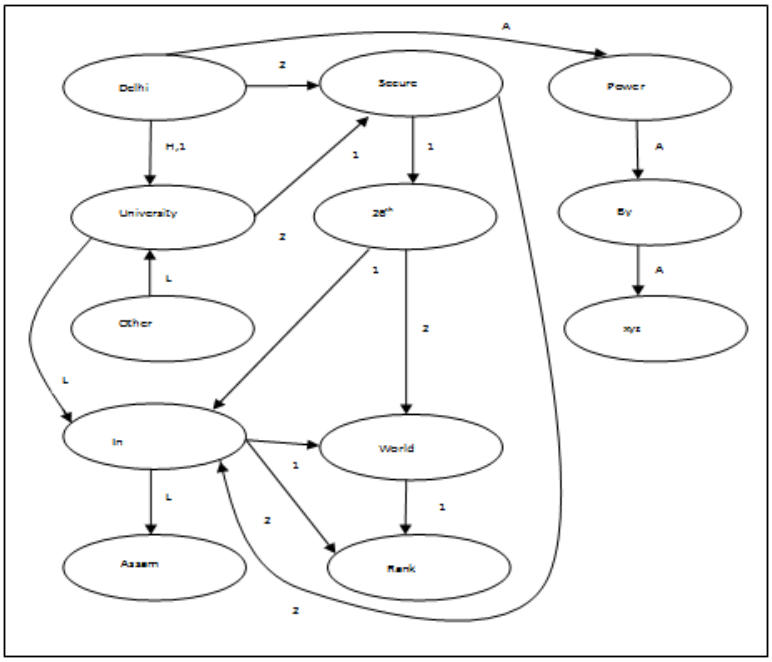

Fig 4: A Web Page Represented by the Composite Model (G_2)

This document is almost similar to the previous web document that we have used to explain the proposed composite model. The document represented by this example has the title "Delhi University", a link whose text reads "Other Universities In Assam", an address that contain "Powered by xyz" and text containing "Delhi University Secures 26th In world Ranking". From fig-3 (G_1) and fig-4 (G_2) we get $max(|G_1|,|G_2|)=13$, $|mcs(G_1,G_2)|= 4$, $\sum d^{\pm}(mcs(G_1,G_2))=17$ $\max(\sum d^{\pm}(G_1),(\sum d^{\pm}(G_2))=38$. By using the prevalent MCS method we get $dist_{MCS}(G_1,G_2)=0.69231$ indicating two dissimilar pages. By using the proposed distance measure we get $dist_{MCS}(G_1,G_2)=0.55263$ which is far better than the previous one.

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.

References

[1] A. Schenker,H. Bunke,M. Last,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),* **(2001)**, pp. 51–54.

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

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

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

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

[7] Khaled 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] Reuters, "Dying for Information: An Investigation into the Effects of Information Overload in the USA and Worldwide", *Based on research conducted by Benchmark Research. London: Reuters Limited,* **(1996)**

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

[10] Horst Bunke a, Kim Shearer , " A graph distance metric based on the maximal common subgraph" , *Pattern Recognition Letters 19* **(1998)**. *255–259*