A Significance-Based Graph Model for Clustering Web Documents

Argyris Kalogeratos and Aristidis Likas

Department of Computer Science,
University of Ioannina,
GR 45110, Ioannina, Greece
{akaloger, arly}@cs.uoi.gr

Abstract. Traditional document clustering techniques rely on single-term analysis, such as the widely used Vector Space Model. However, recent approaches have emerged that are based on Graph Models and provide a more detailed description of document properties. In this work we present a novel Significance-based Graph Model for Web documents that introduces a sophisticated graph weighting method, based on significance evaluation of graph elements. We also define an associated similarity measure based on the maximum common subgraph between the graphs of the corresponding web documents. Experimental results on artificial and real document collections using well-known clustering algorithms indicate the effectiveness of the proposed approach.

1 Introduction

The problem of web document clustering belongs to Web Content Mining area [1] and its general objective is to automatically segregate documents into groups called clusters, in a way that each group ideally represents a different topic. In order to perform clustering of Web documents two main issues must be addressed. The first is the definition of a representation model for Web documents along with a measure quantifying the similarity between two Web document models. The second concerns the employment of a clustering algorithm that will take as input the similarity matrix for the pairs of documents and will provide the final partitioning. Although single-term analysis is a simplified approach, the Vector Space Model is still in wide use today. However, new approaches are emerging based on *graph representations* of documents which may be either *term-based* [1] or *path-based* [2]. The model we propose in this work utilizes term-based document representatives of adjustable size and achieves great modeling performance, while conforming to computational effort conditions (CPU, memory, time).

2 Significance-Based Graph Representation of Web Documents

At first an analysis task is performed to locate the 'useful' information in Web documents, which are primarily HTML documents using a set of tags to designate

G. Antoniou et al. (Eds.): SETN 2006, LNAI 3955, pp. 516 – 519, 2006.

different document parts, and thus assign layout or structural properties. An appropriate model should exploit this information to assign importance levels to different document parts, based on a predefined correspondence between HTML tags and significance levels. In our implementation four significance levels were used: {VERY HIGH, HIGH, MEDIUM, LOW}. Examples of document parts with very high significance are the title and metadata. High significance is assigned to section titles, medium to emphasized parts, and finally the lowest level is assigned to the remainder of normal text.

We represent a document as a directed acyclic graph, well known as *DIG* (*Directed Indexed Graph*), along with a weighting scheme. Formally, a document $d = \{W, E, S\}$ consists of three sets of elements: a set of graph nodes $W = \{w_{1,\ ...,}\ w_{|W|}\}$ each of them uniquely represents a word of the document (unique node label in graph), a set of graph edges $E = \{e_{1,\ ...,}\ e_{|E|}\}$, where $e_i = (w_{k,}\ w_l)$ is an ordered pair (directed edge) of graph nodes denoting the sequential occurrence of two terms in a document. Indeed, we call w_l *neighbor* of w_k and the *neighborhood* of w_k is the set of all the neighbors of w_k. These properties capture semantic correlations between terms. Finally, S is a function which assigns real numbers as significance weights to the *DIG* nodes and edges.

The simplest weighting scheme is actually a non-weighting scheme (*NWM*) [1]. The next step is the assignment of frequencies as graph weights for nodes (*FM*), whereas in this work we propose a more sophisticated significance-based weighting scheme (*SM*). We define the node (term) significance $g_w(w, d)$ as the sum of significance level of all occurrences of w in document d (possible values of significance level of i-th occurrence of w are {VERY HIGH, HIGH, MEDIUM, LOW}).

Regarding to the edges, we should keep in mind the key role they have for document's meaning content, since they represent term associations. Thus, we define the edge significance g_e as:

$$g_e(e(w_k, w_l), d) = \frac{g_w(w_k, d) \cdot g_w(w_l, d)}{g_w(w_k, d) + g_w(w_l, d)} \cdot freq(e(w_k, w_l), d),$$

where $e(w_k, w_l)$ is a document edge and $freq(e(w_k, w_l), d)$ is the edge's frequency in document d. We are now in a position to define the *document content*, which would be based on the weights of all elements of the document graph:

$$g_D^{(all)}(d) = \sum_{j=1}^{nodenum(d)} g_w(w_j, d) + \sum_{i=1}^{edgenum(d)} g_e(e_i(w_k, w_l), d),$$

where $nodenum(d)$ and $edgenum(d)$ are the number of different words and edges respectively in document d. Having estimated the significance values for all elements of the full document graph, we can simply apply a *filtering procedure* on the modeled dataset to keep the P more important nodes per graph. The evaluation criterion can be based either on the frequency weight of a term resulting in a Frequency Filtering (*FF*), or on the significance weight resulting in the proposed Significance Filtering approach (*SF*).

3 Similarity Measure

The next step is to define a measure $s(G_x, G_y)$ that quantifies the similarity between two given document graphs G_x, G_y. This can be enabled through a *graph matching process*, that is based on the maximum common sub-graph between the graphs of the corresponding web documents. The exact computation divides the size of $|mcs(G_x, G_y)|$ of filtered graphs by the $\max(|G_{dx}|, |G_{dy}|)$ of respective unfiltered graphs (note: the size of a graph $|G| = |W| + |E|$). Even though the *mcs* problem is *NP*-complete in general, in our case we have unique graph labels, therefore we deal a reasonable cost of $O(P)$, where P is the global filtering threshold for all documents. This similarity is called graph-theoretical and is used by *NWM*.

In fact, *mcs* ignores whatever information about element significances, even frequencies. We propose the *maximum common content* similarity measure that is based on the significance evaluation of common sub-graphs and is used in combination with the *SM*. In particular, we define two elementary similarity cases:

1. $E_w(w_i^{(x)}, w_j^{(y)}) = g_w(w_i, d_x) + g_w(w_j, d_y)$, which measures the similarity that derives from the mutual word $w_i = w_j$, where $w_i \in d_x$ and $w_i \in d_y$
2. $E_e(e_k^{(x)}(w_i, w_p), e_l^{(y)}(w_j, w_q)) = g_e(e_k(w_i, w_p), d_x) + g_e(e_l(w_j, w_q), d_y)$, which measures the similarity that derives from the mutual edge $e_k^{(x)} = e_l^{(y)}$, where $w_i = w_j$, $w_p = w_q$, $e_k \in d_x$ and $e_l \in d_y$.

If we could define the content union of two documents (at the full graph scale), we could also compute the percentage of common content. Supposing that the *mcs* has been calculated, we evaluate the overall normalized similarity matched sub-graphs:

$$s(G_x, G_y) = \frac{\sum_{i,j,k,l} \left(E_w(w_i^{(x)}, w_j^{(y)}) + E_e(e_k^{(x)}(w_i, w_p), e_l^{(y)}(w_j, w_q)) \right)}{g_D^{(all)}(d_x) + g_D^{(all)}(d_y)} .$$

4 Experiments and Conclusions

We conducted a series of experiments comparing the *NWM* model with the *SM* model proposed in this work. *NWM* uses frequency filtering (*FF*) and assigns no graph weights. The introduced novel *SM* model, on the other hand, uses term filtering based on significance (*SF*) and assigns significance-based weights to graph elements.

As clustering methods, we used an agglomerative algorithm (*HAC*) and two versions of *k-means* algorithm: the typical random center initialization (*RI-KM*) and the *global k-means* (*Global-KM*) [4], already been used to cluster web documents [3].

In our experiments, we evaluate clustering performance using three indices. The first index is the *Rand Index* (*RI*), which is a clustering accuracy measure focused on the pairwise correctness of the result. The second index is a *statistic index* (*SI*), which computes the percentage of N documents assigned to the "right" cluster, based on ground truth information. A third index we considered is the typical *Mean intra-Cluster Error* (*MCE*). Three web document collections were used: the *F-series*

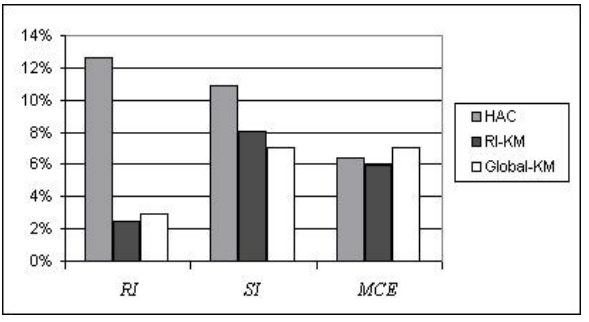

Fig. 1. *SM* vs *NWM* overall *improvement* on all collections using three indices

(95 web documents from 4 classes) and *J-series* (185 web documents from 10 classes) used in [7] and an artificially created dataset consisting of classes of high purity.

The experimental results (Fig. 1) indicate the overall improvement obtained using the proposed *SM* approach. We have found that *SM* is superior to *NWM* in all cases since a clear improvement for all indices was observed in almost all experiments. In what concerns the clustering algorithms, the agglomerative approach exhibits sensitivity on "difficult" data, while when used with the *SM* model, it can be competitive to k-means type of algorithms. From the k-means class of methods, *Global-KM* shows a clear qualitative superiority comparing to *RI-KM,* which nevertheless also remains a reliable and computationally "cheap" approach.

References

1. A. Schenker, M. Last, H. Bunke and A. Kandel: Clustering of Web Documents Using a Graph Model, Web Document Analysis: Chalenges and Opportunities, eds. A. Antonacopoulos and J. Hu, to appear
2. K. M. Hammuda: Efficient Phrase-Based Document Indexing for Web-Document Clustering, IEEE, 2003
3. A.Schenker, M.Last, H. Bunke, A.Kandel: A Comparison of Two Novel Algorithms for Clustering Web Documents, 2nd Int. Workshop of Web Document Analysis, WDA 2003, Edinburgh, UK, August 2003
4. A. Likas, N. Vlassis and J. J. Verbeek: The global k-means clustering algorithm, Pattern Recognition, Vol. 36, 2003, pp. 451 – 461