Concept Graph Preserving Semantic Relationship for Biomedical Text Categorization

Chetna Gulrandhe
Dept. of Computer Science & Engg.
WCEM, Nagpur, India
chetna298@gmail.com

Chetan Bawankar
Dept. of Computer Science & Engg.
WCEM, Nagpur, India
chetan251htc@gmail.com

Abstract

Recently, graph representations of text have been showing improved performance over conventional bag-of-words representations in text categorization applications. In this paper, we present a graph-based representation for biomedical articles and use graph kernels to classify those articles into high-level categories. In our representation, common biomedical concepts and semantic relationships are identified with the help of an existing ontology and are used to build a rich graph structure that provides a consistent feature set and preserves additional semantic information that could improve a classifier's performance. We attempt to classify the graphs using both a set-based graph kernel that is capable of dealing with the disconnected nature of the graphs and a simple linear kernel. Finally, we report the results comparing the classification performance of the kernel classifiers to common text based classifiers.

Keywords: Text categorization, text retrieval, query processing, graph kernels, graph representations, biomedical ontology's, mining methods and algorithms, text mining, classifier design and evaluation, modelling structured, textual and multimedia data

1. Introduction

Biomedical electronic document databases are growing exponentially, resulting in huge digital repositories. Organizing and searching these documents manually is increasingly costly and time consuming. MedLine is one example of a fast growing biomedical digital library. It currently has more than 18 million indexed articles and therefore its availability and usability has become critical to students and researchers working on biomedical-related topics. With the rapid growth, biomedical literature has been the subject of intensive information retrieval and machine learning investigations throughout past decades. Text categorization (also known as document categorization) is a challenging research area where text documents are categorized using predefined labels based on their content. Applying improved text categorization techniques to the biomedical databases is essential to overcome the information overload problem and to facilitate indexing, filtering and managing the growing number of articles in those databases.

Most of the existing text categorization techniques use a vector representation of documents. In the vector space model, key entities and concepts are identified from text and used as features. The disadvantage of the vector representation is the lack of semantic relationships among key entities and concepts in the text. Recently, graph mining and graph modelling techniques have begun to gain popularity in modelling complex data such as protein sequences and structures and social networks. The advantage of graph modelling is the use of "rich" semantic representation of relationships among key entities and concepts in a text and hence may yield improved results when classifying documents. In addition, kernel functions for graphs and other structured data have garnered particular interest. Kernel functions are an elegant method of embedding non-vector data, such as graphs, into a vector space suitable for operations using existing classifiers.

Text categorization (also known as text classification or topic spotting) is the task of automatically sorting a set of documents into categories from a predefined set. This task has several applications, including automated indexing of scientific articles according to predefined thesauri of technical terms, filing patents into patent directories, selective dissemination of information to information consumers, automated population of hierarchical catalogues of Web resources, spam filtering, identification of document genre, authorship attribution, survey coding, and even automated essay grading. Automated text classification is attractive because it frees organizations from the need of manually organizing document bases, which can be too expensive, or simply not feasible given the time constraints of the application or the number of documents involved. Text categorization has recently become an active research topic in the area of information retrieval. The objective of text categorization is to assign entries from a set of prespecified categories to a document. A document here refers to a piece of text. Categories may be derived from a sparse classification scheme or from a large collection of very specific content identifiers. Categories may be expressed numerically or as phrases and individual words. Traditionally this categorization task is performed manually by domain experts. Each incoming document is read and comprehended by the expert and then it is assigned a number of categories chosen from the set of prespecified categories. It is inevitable that a large amount of manual effort is required. For instance, the MEDLINE corpus, which consists of medical journal articles, requires considerable human resources to carry out categorization using a set of Medical Subject Headings (MeSH) categories.

Text Categorization may be formalized as the task of approximating the unknown target functions Ø: D*Č—›{T,F} (that describes how documents ought to be classified, according to a supposedly authoritative expert) by means of a function Ǿ: D * Č —› {T, F} called the classifier where Č={C1 . . ., C|c|} is a predefined set of categories and D is a (possibly infinite) set of documents. If Ø (dj, ci) = T, then dj is called a positive example (or a member) of ci, while if Ø (dj, ci) = F it is called a negative example of ci.

2. Related Work

Several supervised learning techniques have been proposed to automate the manual process of classifying documents. Those include NB classification, SVMs, k-NN classification, and Decision Trees. Graph representations have also been used to categorize documents based on graph matching where the complex structure of documents can be represented as nodes and edges that encode the textual features of the documents. The addition of relationship edges to describe documents can create a much higher-dimensional feature space, thus allowing for more nuanced and potentially useful embeddings of the documents. Weighted frequent subgraphs were used in to construct effective feature vectors for classification and to overcome the computation overhead that is associated with graph structures. Arey and Chakravarthy used exact and inexact graph matching as well as substructure pruning and ranking to optimize classification and compare their result to a naive Bayesian classifier. Gee and Cook exploited the linguistic syntactic and semantic characteristics of phrases in text. They encoded phrases as graphs and used a substructure and pattern discovery algorithm for classification. The relationships used to connect graph nodes can be as diverse as the applications. Chen et al. proposed a graph representation for document summarization tasks. They used a thesaurus and association rules to connect key phrases in the text. Wan et al. also used graphs to represent documents for summarization. They use three graphs to capture word-word, word-sentence, and sentence-sentence relationships in the text and compute word and sentence saliency scores to rank their results.

A common preprocessing used for graph classification is projecting the graph onto a kernel space using a kernel function. One possible kernel function can be defined as an inner product between two graphs and must be positive semi-definite and symmetric. Such a function embeds graphs or any other objects into a Hilbert space, and is termed a Mercer kernel from Mercer's theorem. Kernel functions can enhance classification in two ways: First, by mapping vector objects into higher dimensional spaces; second, by embedding non-vector objects in an implicitly defined space.

3. Proposed Method

This method consists of two major components. The first is the graph construction part, which involves mapping biomedical terms that are extracted from the text into predefined concepts of a controlled vocabulary. In addition, the relationships among the concepts are also identified and added to the representation. The second component is the application of a graph kernel function to compute the similarities between the generated graphs and a kernel classifier to discriminate between the documents given their embedding in the kernel space.

Fig. 1 shows the data flow of the procedure of extracting concepts and relationships as well as feeding them into a graph kernel function for classification. In brief, the process is as follows: First, a set of biomedical articles are selected from different journals; next, biomedical concepts are extracted from the documents and mapped to concepts from the UMLS database; concept relationships are then extracted and used to link the concepts, resulting in the concept graphs; a kernel matrix is prepared by computing similarities between the graphs; and finally, the kernel matrix is used for learning and prediction of the documents' target classes. The overall process consists of two phases: 1) graph construction and 2) classifier learning and output. Each phase is described in detail in the following sections.

3.1 Graph Construction

The graph construction phase begins by collecting a set of published articles from different journals. The articles were grouped by the journal in which they were published. The journals represent high-level categories of biomedical related disciplines and, thus, are used as the class labels for the different sets of documents. The text content is then used to construct a set of concept graphs, where each document is represented by one graph. Several keywords were chosen as class labels for the graphs to be constructed and were used to query the Medline database for articles that contain those keywords in both their title and abstract. The keywords are biomedical terms that represent a general topic (ex: spinal cord injury) or a common biomedical entity name (ex: insulin).

To extract the concepts from the text, all noun phrases are first identified using a part-of-speech (POS) tagger and regular expressions. The POS tagger labels all lexical items of each sentence, and the regular expressions are used to identify common patterns of the items in a sentence that make up a noun phrase. At this stage, all noun phrases, such as Idiopathic Scoliosis or Kidneys, are considered potential concept candidates to be added to the graph representation.

To ensure the target concepts correspond to a controlled vocabulary set, we then attempt to map the n-grams of each noun phrase into biomedical concepts of the UMLS database. If any of the n-gram substrings is found in UMLS, it is added to the corresponding graph as a concept node and each assigned a unique identifier. A concept string in UMLS might refer to multiple concepts with different meanings whereas a concept unique identifier (CUI) refers to only one concept associated with one or more string descriptors that might slightly vary because of the different vocabulary sources in UMLS.

Concepts that have the same string descriptors but different meanings are implicitly disambiguated by the weighting technique described in the following section, which favours nodes that indicate more connectivity in a graph. Mapping the terms into predefined concepts also allows us to look for possible relationships among them within UMLS. For each pair of nodes, we attempt to find a relationship in UMLS and add it as an edge between the nodes if it exists. The available relationships are of semantic nature some of which are synonym, parent-child, and sibling relationships. Fig. 2 shows a sample text and the corresponding concept graph with the extracted nodes and edges. It is worth noting here that we do not explicitly use the specific types of the relationships between concepts. An edge is added to the graph whenever the corresponding concepts are related, regardless of what type of relationship exists between them.

3.2 Node and Edge Weights

All nodes in the graph are consequently assigned four different weight components that correspond to their significance in a document. Below is a description of each:

- $F_{i,d}$: Concept frequency, which is the number of times a concept term i appears in a document d. This value assigns more weight to concept terms with high occurrence frequency in a document.
- idf_i: Inverse frequency of documents that contain a concept term i. This value ensures that common terms in the whole data set are given lower weights while rare terms are favoured.
- cw_i: Connectivity weight of a concept node i in a graph. This is the calculated as the magnitude of the vector of f * idf values of related nodes c1; c2; . . . ; cj. This component assigns higher weight values to concept nodes that are better connected in a graph. Nodes that are connected to more nodes of high f * idf values would be favoured.
- cs_i: Cluster size, which is the number of nodes of the cluster containing the concept node i in a graph. In this experiment, clusters are referred to as all connected components of the containing graph.

All values are then normalized using min-max normalization, and the product of the weight components is calculated for each concept node i in a document d as such:

$$NW_{i,d} = F_{i,d} * idf_i * cw_i * cs_i$$

The related nodes' weights are aggregated into a single value and assigned to the corresponding edges. The weight of an edge k is, thus, calculated as the sum of weights of its terminal nodes i and j in a document d:

$$EW_{k,d} = NW_{i,d} + NW_{j,d}$$

To reduce the dimensionality of the feature space, edges saving weights below a certain threshold were dropped from the feature set. Although the threshold used was very low, the number of unique edges was significantly reduced to around 10 percent of the original numbers, as most of the extracted edges are not significant and not representative of the documents.

3.3 Classifier Learning and output

After transforming the set of articles into a set of graphs, a graph kernel function is applied to compute the similarity between all pairs of graphs, and the resulting kernel matrix is used for classification. The first is a simple set-based kernel that is used to measure concept graph similarity based on the number of shared edges.

There are a couple properties that make a set based kernel function attractive. The first reason is that the set computations used are easily implemented and understood, leading to a kernel function that is easy to interpret, which results in a greater confidence in producing reliable measures of graph similarity. The second reason is that many of the concept graphs are disconnected or sparse, with many more nodes than edges, which can pose problems for some graph mining algorithms. This kernel function is based on the Jaccard coefficient. It computes the similarity between two graphs X and Y as the ratio of the cardinality of the intersection of the edges sets Ex and Ey to the cardinality of their union:

$$K(x,y)= \frac{|E_x \mid E_y|}{|E_x \, U \, E_y|}$$

The second is a common normalized linear kernel based on the cosine similarity between the edge weight vectors of a pair of graphs. The kernel function returns a normalized inner product of the weight vectors:

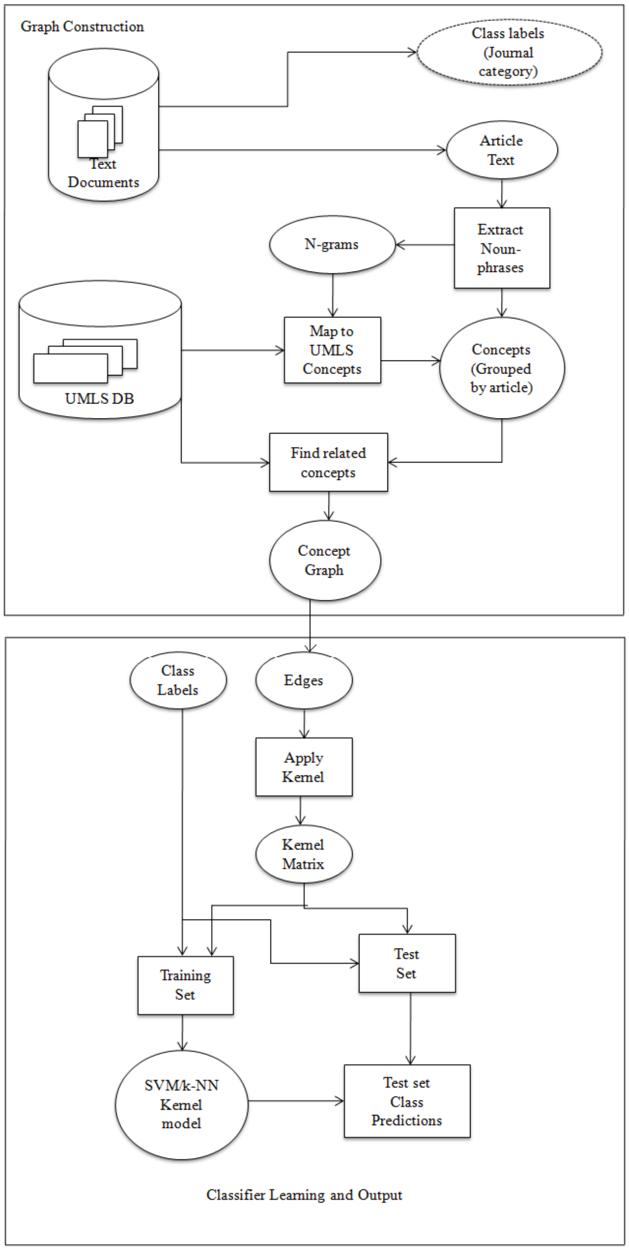

Fig 1: System Overview

$$K(x,y)= \frac{|W_x \mid W_y|}{||W_x|| \; || W_y||}$$

Once a kernel between all graphs is computed, the graphs' similarities result in a kernel matrix. This matrix can then be used in a kernel-based classifier to make predictions on new data. We used the kernel matrix with a SVMs classifier and a k-NN classifier to make classification predictions, or in other words, to predict to which journal a certain document belongs.

Example:

The presentation is provided, concerning the medical history, clinical examination, conventional radiography, stereo-radiography, surface topography, ultrasounds, computed tomography, and magnetic resource imaging, focusing on the points specific for the pathology of idiopathic scoliosis. Use of the scoiimeter became systematic in the clinical evaluation. Quality of live Questionnaire, including those endorsed by the Society f Scoliosis Orthopaedic and Rehabilitation Treatment (SOSRT), oriented towards scoliotic patients, again on popularity and are extremely helpful to objectively evaluate the disability elated to scoliosis. Classical radiography serves as the basic exam to determine the curve type and magnitude. Ultrasounds, computed tomography and magnetic resonance imaging are indicated in precisely defined clinical situations. Stereo-radiography and surface topography seem to be the most promising techniques, however requiring standardization. Apart from sophisticated measurements, the experience of a physician cannot be underestimated. High standard clinical evaluation will probably continue to serve as a reference for other methods of assessment of patients with scoliosis.

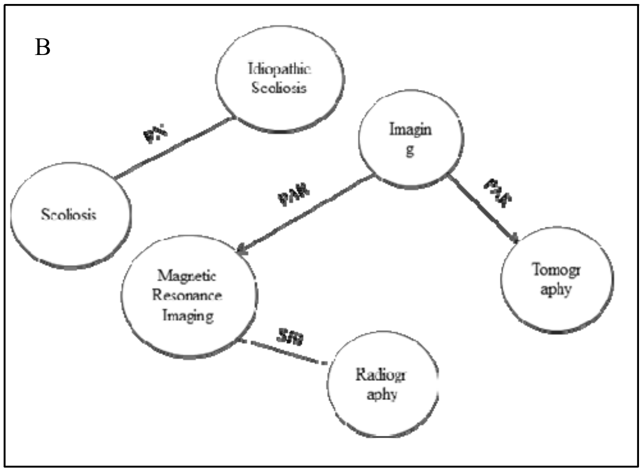

Fig 2: Sample Text and Corresponding Graph

4. Conclusion

Categorizing biomedical text is a challenging problem due to the huge number of articles published every year. In this study, we propose a promising approach to text categorization based on building concept graphs to represent documents and classifying them using a k-NN classifier. The results show that the rich representation of documents, whereby related biomedical concepts are added to the model, significantly improves the classification accuracy. It is interesting to note here that in some cases the added information (related concepts) didn't contribute positively to the classification until the semantic relationships (edges of the graphs) were used.

However, the statistical significance of the improvement using semantic relationships is very strong. We believe that using a well-trained NER module and a more accurate concept identification technique will lead to even greater improvements. SVMs have shown great results in classification as well and are also worth trying with our technique.

References

[1] Minakshi Mishra, Jun Huan and Min Song, "Text Categorization of Biomedical Data Sets using Graph Kernel and Controlled Vocabulary" IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 10, NO. 5, SEPTEMBER/OCTOBER 2013, pp 1211-1217.

[2] S. Bleik, M. Song, A. Smalter, J. Huan, and G. Lushington, "CGM: A Biomedical Text Categorization Approach Using Concept Graph Mining," Proc. IEEE Int'l Conf. Bioinformatics and Biomedicine Workshop (BIBMW '09), pp. 38-43, 2009.

[3] Fabrizio Costa and Bjorn Bringmann, "Towards Combining Structured Pattern Mining and Graph Kernels" IEEE International Conference on Data Mining Workshops, pp. 192-201 2008.

[4] R. Angelova and G. Weikum, "Graph-Based Text Classification: Learn from Your Neighbors," Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 485-492, 2006.

[5] M. Mishra, J. Huan, S. Bleik, and M. Song, "Biomedical Text Categorization with Concept Graph Representations Using a Controlled Vocabulary," Proc. 11th Int'l Workshop Data Mining in Bioinformatics, pp. 26-32, 2012.

[6] B.V. Dasarathy, "Nearest neighbor (NN) norms: NN pattern classification techniques", *IEEE Computer Society Press*, Los Akunitos, California, 1991.

[7] K.M. Borgwardt and H.P. Kriegel, "Shortest-path kernels on graphs", *Proceedings of the International Conference on Data Mining (ICDM)*, 2005.

[8] A. Wilcox, G. Hripcsak, and C. Friedman, "Using Domain Knowledge Sources to Improve Classification of Text Medical Reports*", Proceedings of ACM SIGKDD Workshop on Text Mining*, 2000.