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. The relationships used to connect graph nodes can be as diverse as the applications. Word and sentence saliency scores to rank their results.

A kernel function is a mapping between a pair of graphs into a real number. This function defines 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. The advantages of mapping objects into a higher dimensional space, the so called kernel trick, are apparent in a variety of cases where objects are not separable by a linear decision boundary. This implicit embedding is not only useful for non-linear mappings, but also serves to decouple the object representation from thespatial embedding. A kernel function need only be defined between data objects in order to apply a kernel classifier. Such a kernel classifier can then be used for classification of graph objects by defining a kernel function between graphs, without explicitly defining any set of graph features.

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 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.

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.

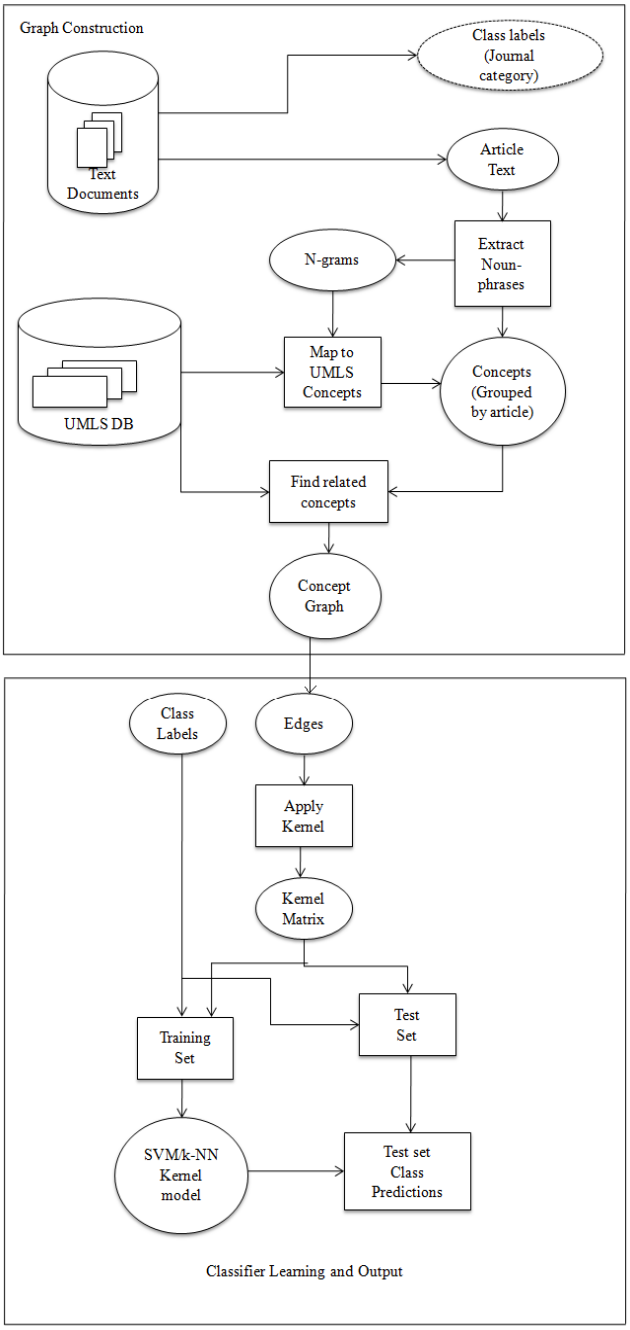

Figure 1: System Overview

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:

$$K(x,y)= \frac{|Ex \mid Ey|}{|Ex \cup Ey|}$$

3.4 The Algorithm

The above discussed technique can be implemented as follows:

Step 1: 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).

Step 2: The titles of the retrieved articles are then used as graph labels and the abstracts are passed into a named entity recognition module. Only abstracts with one or more a named entity recognition (NER) module and a concept identification module. The second is the application of a graph kernel function to compute the similarity between the generated graphs and a kernel classifier to discriminate between papers given their embedding in the kernel space.

Step 3: For each article, the entities are used to query the UMLS database and are subsequently mapped to predefined concepts. For each entity within the article, the top three concepts (which are assumed to be the most relevant) are selected and added to the corresponding article's concepts set.

Step 4: Now that the graphs consist of well defined UMLS concepts, the UMLS database is queried to find additional related concepts. The relations are already defined in the database with labels describing the nature of those relations. We then try to retrieve concepts having a "parent-child" or "synonym" relationship with the existing concepts, add those to the graphs, and add the relations as edges between the nodes with the corresponding label.

Step 5: The mapping of node concept labels to integers is more complicated because these concepts are often long strings containing a number of different words. There are a large number of unique concepts and similar concepts do not always have the exact same words/text in the same order within them. Therefore similar concepts must be grouped and then all concepts in a group are mapped to the same integer label.

Step 6: The concept strings are decomposed into a "bag of words" representation and then grouped according to the number of shared words.

Step 7: Concepts that share a large number of the same words are grouped and mapped to the same integer label. This process is carried out without any knowledge.

Step 8: After transforming a set of papers into a set of graphs, a graph kernel function is applied to compute the

similarity between all pairs of paper graphs, and the resulting kernel matrix is used for classification.

4. Result and Analysis

The datasets are comprised from various journals on radiology. We obtained our training and test data sets from the kernelmatrix and the documents' class labels. In each validation trial, one set was reserved for testing and the others were used for training.

Data Set	No. of Samples
Radiology	20
Cancer	22
Neurology	15

The accuracy is calculated using Precision and recall algorithm. For classification tasks, the terms true positives, true negatives, false positives, and false negatives (see also Type I and type II errors) compare the results of the classifier under test with trusted external judgments. The terms positive and negative refer to the classifier's prediction (sometimes known as the expectation), and the terms true and false refer to whether that prediction corresponds to the external judgment (sometimes known as the observation).
Precision and recall are then defined as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Recall in this context is also referred to as the true positive rate or sensitivity, and precision is also referred to as positive predictive value (PPV); other related measures used in classification include true negative rate and accuracy. True negative rate is also called specificity.

$$\text{True Negative Rate} = \frac{TN}{TN + FP}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Data Set	Precision	Recall	Accuracy
Radiology	0.844	0.831	0.853
Cancer	0.844	0.864	0.849
Neurology	0.927	0.923	0.925

5. 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 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, pp 1211-1217, 2013.
[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.