Learning Text Classifier using the Domain Concept Hierarchy

Bill B. Wang, R *I.* (Bob) *McKay, Hussein A. Abbass, Michael Barlow*
School of Computer Science, University College, ADFA, University of New South Wales
Canberra, ACT 2600
Email: {biaowang, rim, h.abbass, spike}@cs.adfa.edu.au

Abstract: Automatic text categorization is an important component in many information organization and management tasks. Research has shown that similarity based categorization algorithms like K-nearest neighbour (KNN) are effective in document categorization. These algorithms use index terms to represent documents. However some drawbacks persecute these algorithms. One major drawback is that they tend to use all features when computing the similarities, which implies that they must search in a high-dimensional space. Another major drawback is that they tend to use a very large training document set so that all terms, which are important to identify content of documents, are covered. To overcome these drawbacks, in this paper, we present a novel method to search for the optimal representation in a domain ontology hierarchical structure to reflect concepts for the taxonomic standard for pre-defined categories. Experiments have shown this is a feasible method to reduce the dimensionality of the document vector space effectively and reasonably and consequently improves the generalisation power of the derived classifier. The result is a classification method which is both very significantly less costly, in computation terms, and yet of considerably higher accuracy than comparable methods.

Keywords: text classification, ontology, concept hierarchy, KNN algorithm.

I. Introduction

Automatic text categorization [1] is the task of assigning natural language texts to one or more pre-defined categories based on their content. As the volume of text documents available on the Internet, digital libraries and corporate intranets increase, text categorization is increasingly important to help people find information from these vast resources. Text categorization presents huge challenges due to a large number of features, feature dependency, multi-modality and large training sets.

A growing number of statistical learning methods have been applied to this problem in recent years, including Bayes belief networks, decision trees, support vector machines, neural networks and K-nearest neighbour (KNN) classifiers. These methods all use index terms to represent documents. Many researches have shown that similarity based categorization algorithms such as KNN and centroid based classification are very effective in document categorization [3]. A cross-experiment comparison [2] between 14 major categorization methods, including KNN, decision tree, naive Bayes, linear least squares fit, neural network, SWAP-1, Rocchio , etc., has shown that KNN is the one of top performers and it performs well in scaling up to very large and noisy categorization problems. However, these effective categorization algorithms still suffer from major drawbacks that greatly limit their practical performance.

One major drawback of these algorithms is that they see all words as potential features for a document, implying that they compute the similarity between documents in a high-dimensional space. Empirical and mathematical analysis [4] [5], however, has shown that finding the nearest neighbours in high-dimensional space is very difficult because most points in high-dimensional space tend to have equal distances from all the other points. In fact, in many document data sets, only a relatively small number of the total features may be useful in categorizing documents, and using all the features may affect performance. So determining how to reduce the length of document vectors effectively and reasonably is a challenge for categorization researchers. Stop words lists and word stemming are some of the earliest effort in this problem. In recent years, many term-weighting algorithms [3] have been developed to extract features from documents. However, all are statistics-based, in other words, either word- or syntax-based.

Another major drawback is that these algorithms need a large training document set covering all terms, which are important to identify content of documents. The KNN classifier is an instance-based classifier, which means an ideal training document set for one particular category will cover all important words and possible distribution in this category. In other words, a text that uses only some key words out of a training set may be assigned to the wrong category. In practice, however, establishing such a training set is infeasible or impossible.

This paper presents a novel method to find an optimal concept representation by searching a domain-specific concept ontology for a particular text classifier (KNN in this paper) to overcome the above drawbacks effectively and reasonably. Obviously, different training sets have their own taxonomic standards. This means they use different concept levels to identify document content. For example, the names of different kinds of heart disease are key index terms to identify text content if each category represents a kind of heart disease. However, all these names can be mapped to the concept "heart disease" to identify text content if each category is one of the 23 major MeSH (Medical Subject Headings [11]) diseases categories. The principal idea of our approach is to establish a hierarchical concept structure based on

domain-specific concept ontology for a particular training document set, and then to use a heuristic search algorithm to search into this hierarchical structure to find an optimal concept representation. This will be one which maximizes the difference, between the distances between documents belonging to the same category, and the distances between documents belonging to different categories. In addition to effective and reasonable reduction of dimensionality of vector space, using higher level concepts to represent documents may assign a text, which uses some new key words out of the training set, to a correct category provided these new terms can be mapped to concepts in the ontology structure.

This paper is structured as follows: section 2 briefly introduces the notion of domain-specific concept ontology and UMLS knowledge resources, section 3 describes the process of this system, some experimental results and discussions are presented in section 4, finally the conclusion is given in section 5.

II. Domain-specific concept ontology and UMLS knowledge resources

The term ontology represents various meanings when it is used in different ways and in different disciplines. Despite these differences, computer scientists and metaphysicians use the term ontology to describe formal descriptions of objects in the world, the properties of those objects, and the relationships among them. In artificial intelligence, according to Gruber [6] an ontology is a specification of a conceptualization. It defines the vocabulary of a domain and constraints on the use of terms in the vocabulary.

In our research, domain-specific concept ontology specifies the terms that are used to represent documents, the categories attached to these terms, and the relations (ISA in this paper) which exist between terms and categories (Figure 1). The hierarchical concept structure, which we use for a particular training document set, is a part of domain-specific concept ontology based on terms used in the training set. The process to establish this structure is introduced in section 3.

I. Domain-Specific Concept

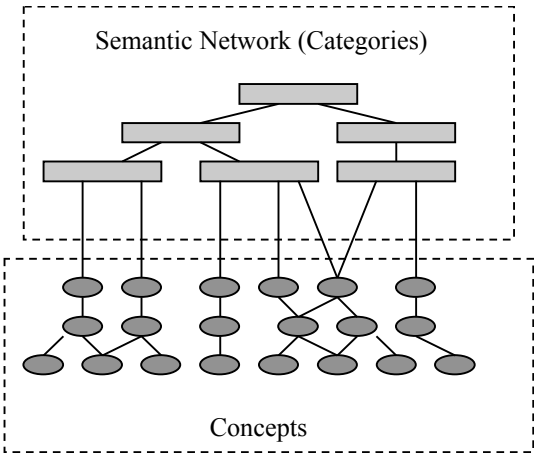

Fig 1 . A Draft of Domain-Specific Concept Ontology

Obviously, our method has a limitation that it is only able to apply in the domains that have fully developed domain ontoloiges. There are few domain ontologies available now but many domain ontologies are under construction in the development of the semantic Web.

The Unified Medical Language System (UMLS) [7], a set of knowledge sources developed by the US National Library of Medicine, can be viewed as a complete concept ontology for medical domains. It consists of three sections: a metathesaurus, a semantic network and a specialist lexicon, and contains information about medical terms and their inter-relationships. It is organised by concept, and contains over 800, 000 concepts and 1.9 million entries. Various types of relationships between concepts are defined in this system. ISA is the primary relationship. We establish the hierarchical concept structure based on this system for a particular training set, which contain documents in medical domains.

III. Establishing Concept Representation

There are four major steps to establish concept representation for documents.

1. Map document terms to concepts based on UMLS
2. Establish a concept hierarchy for the document set
3. Use hill-climbing to search the concept hierarchy for the optimal representation based on a fitness function
4. Establish a new description of documents based on the optimal representation found in the previous step

A. Mapping Terms to Concepts

The most straightforward representation of documents relies on term vectors. The major drawback of this basic approach for document representation is the size of the feature vectors, usually more than 10,000 terms. In the application of text categorization, however, completely different terms may represent the same concepts. In some cases, terms with different concepts can even be replaced with only one higher level concept without negative effect on performance of the classifier. For example, ANEMIA and LEUKEMIA can be replaced with the higher level concept HEMATOLOGIC DISEASE in many situations of text categorization. Obviously, mapping terms to concepts is an effective and reasonable method to reduce the dimensionality of the vector space.

The mapping relies on the API provided by UMLS. We use two query functions – left truncation (TL) and exact match (EM) – provided by the UMLS query interface. Left truncation query will back all concepts that start with query terms. Exact match query will execute lexicon analysis by UMLS query system automatically.

We aim to find the maximal concept units in each sentence. For example, consider the sentence

AIDS is a kind of human immunodeficiency virus.

According to the mapping algorithm defined below, we will get two concepts: ‘AIDS’ and ‘HIV’. ‘Human’ and ‘virus’ are not viewed as independent concepts, even though they do occur in the ontology.

Through this mapping process, we will get concept sets for individual documents. Each will include all distinct

concepts and their frequency from the individual document. We will also get a shared concept set for the document set, which includes all distinct concepts from the whole document set.

Term2Concept Mapping Algorithm:

Input: a complete sentence from a document

```
1.  begin
2.     take the first word from the sentence
3.     TL query the current term
4.     if (true) then
5.          take the next word to append
6.          goto 3
7.       else
8.          EM query the current term
9.           if(true) then
10.               put current term into output set
11.               remove the current term from sentence
12.               goto 18
13.            else
14.               if (more than one word) then
15.                  remove last word from current term
16.                  goto 8
17.               else
18.                 if (next word) then goto 2
19.                    else goto end
20. end
```

output: a set of concepts for this sentence

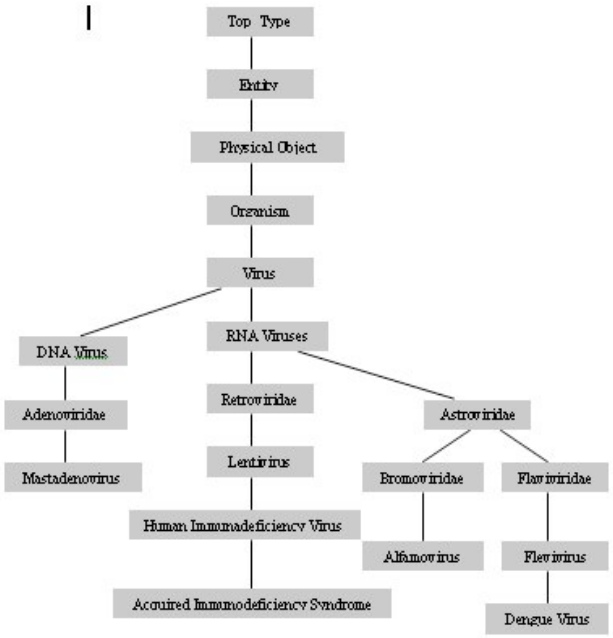

Fig 2 . A Sample Concept Hierarchical Structure

B. Establishing the concept hierarchy

The UMLS query interface provides a parent query function for retrieving parents of concepts. The concept hierarchic structure is established by repeatedly querying parent from shared concepts up to the root of the semantic network. The completed concept hierarchic structure is a fully-connected graph rooted at 'top_type'.

For instance, it is assumed that there are only five distinct concepts as below occurring in a document set.

[*Mastadenovirus, AIDS, Human Immunodeficiency Syndrome, Alfamovirus, Dengue Virus*]

Based on this shared concept set, we will get the concept hierarchic structure in Figure 2. From this structure, we can see that several combinations of different level concepts can be chosen to represent documents for different taxonomic standards, e.g. [Virus], [DNA Virus, RNA Virus] and [DNA Virus, Retroviridae, Astroviridae]. All original concepts can be mapped to these higher level concepts.

C. Seeking optimal representation in the hierarchy

Here, we use the hill climbing algorithm to search the concept hierarchical structure obtained in the previous step to find the optimal representation (a combination of concepts) for a particular document set. Our aim is to maximise the difference between the within-category document-to-document distances, and to maximize the between-category document-to-document distances.

First, we establish a copy of the hierarchical structure for each document, and assign a frequency to each concept node as edge. The edge of a parent concept node is the sum of the edges of all child nodes.

$$C_i = \begin{cases} Ln(1 + tf_i / Maxtf) + 1 & tf_i > 0 \\ 0 & tf_i = 0 \end{cases}$$

The vital problem is to define an appropriate fitness function for the hill climbing search algorithm. The first step is to use the function below to obtain the normalized concept frequency for concept vectors.

C_i is the ith concept of a concept vector, tf is the raw frequency of a concept, $Maxtf$ is the frequency of the most frequent concept in this concept vector.

To allow for variation in document size, the concept frequency is normalized by frequency of the most frequent term in the document [10]. Then, the natural log is taken, so that we measure the ratio of frequencies of terms rather than their absolute frequencies (this could also be given an information-theoretic interpretation).

We define the distance between document i and document j as below.

$$d_{ij} = \sqrt{\sum_{k=0}^{n} (c_{i,k} - c_{j,k})^2}$$

Where n is the length concept vector.

Finally, we define the fitness function for the hill climbing search algorithm as below.

$$fitness = \sum_{i \in D} \frac{\sum_{j \in D_{ex}} d_{ij}}{1 + \alpha \sum_{j \in D_{in}} d_{ij}}$$

D is the set of all documents in training set, D_{ex} is the set of all documents that do not belong to the category which contains document i, and D_{in} is the set of all documents

belonging to the same category as document *i*. So the fitness is proportional to the sum of distances between documents belonging to different categories. It is inversely proportion to the sum of distances between documents belonging to the same categories. The α parameter adjusts the relative importance of numerator and denominator. We suggest that α should be greater than 1 to emphasis the importance of minimization of the distances between documents belonging to the same categories if bottom-up search is used, while α should be less than 1 to emphasis the importance of maximization of the distances between documents belonging to different categories if top-down search is used.

We define the top-down hill climbing search algorithm in our research as below.

Initial status: fitness = 0;
optimal concept set (OTS) = [top_type];
best fitness (BF) = 0;
best concept set (BCS) = ∅;

1. begin
2. take the first concept from OTS
3. if(has child) then
4. create as temporary concept set TCS = OTS
5. use children to replace parent concept in TCS
6. count temporary fitness (TF) with TCS
7. if(TF > BF) then
8. BF = TF
9. BCS = TCS
10. if(next concept of OTS) then
11. take the next concept
12. goto 3
13. else if(BF > fitness) then
14. OTS = BCS
15. fitness = BF
16. goto 2
17. else goto end
18. else take the next concept
19. goto 3
20. end

output: optimal concept set

D. Concept representations for documents

Based on the optimal concept set obtained from last set, we can establish new descriptions for both training documents and test documents.

IV. Experiment Setup and Results

A. Document collection

We chose documents from 5 journals from the 2635 in the MEDLINE database [8] to form our training and test document sets as in table 1. These documents were chosen randomly by people without specialized medical knowledge. We used title plus abstract as text for this experiment. These journals cover non-overlapping categories, hence we have chosen our concept classes to be the subject matter of each journal.

B. Accuracy measure

To evaluate the trained classifier on test documents for each class, an accuracy measure is defined:

$$Accuracy_i = \frac{c_i}{n_i + w_i/2}$$

c_i is the number of correctly classified documents in class *i*, n_i is the number of test documents in class *i*, and w_i is the number of documents that are wrongly assigned to class *i*. In this equation, we consider that the accuracy of a particular category depends on both the number of documents which are correctly assigned to this category, and the number of documents which are wrongly assigned to this category.

Table 1. Number of Training/Test Documents

Journal Name	Category Name	Num of Train	Num of Test
Addiction (Abingdon, England)	Addiction	23	8
AIDS Care	AIDS	23	8
American Heart Journal	Heart	25	9
BMC Cancer	Cancer	26	9
The British Journal of Ophthalmology	Ophthalmol	26	9
Overall		123	43

To evaluate the overall performance of the classifier, we define an accuracy measure as below.

$$overall = \frac{C}{N}$$

C is the number of correctly classified documents, and *N* is the number of test documents.

C. Summary of results

By searching UMLS resources, 1121 distinct concepts were obtained from the 123 training documents. With top-down hill climbing search of the concept hierarchy, this was reduced to 63 high-level concepts. Table 2 shows the results produced by KNN classifier based on the above concept representation. Where K = 5.

Table 2. Experimental results with optimal concepts

Category	Accuracy of Original Train Documents	Accuracy of Test Documents
Addiction	91.7%	66.7%
AIDS	87.5 %	73.7%
Heart	92%	70%
Cancer	100%	84.2%
Ophthalmol	92.6%	63.2%
overall	95%	79%

For the purpose of comparison, we show the results produced by the same KNN classifier with the original 1121 concepts as vector features in table 3.

The results show a dramatic improvement in accuracy even on the training data, in itself perhaps a surprising result. More importantly, the clustering generalises extremely well to the test data, providing a clustering with a perfectly usable performance, though this could not be said of the classification derived from the original 1121 concepts, whose accuracy in the opthalmology and heart classifications is only around double that of random guessing. The test set classification accuracy, derived from only 123 training examples, is in the ballpark that would normally be associated with classifications trained on tens or hundreds of thousands of documents.

Table 3. Experimental results with original concepts

Category	Accuracy of Original Train Documents	Accuracy of Test Documents
Addiction	82.4%	63.2%
AIDS	73.1%	66.7%
Heart	77.8%	38.1
Cancer	80%	57.1%
Ophthalmol	74.1%	40%
overall	83.7%	62.8%

V. Conclusion

Our results show that seeking the optimal concept representation in a hierarchical structure is a viable method to effectively reduce the length of document vectors. The results are good compared with those usually reported for statistics-based term weight algorithms [2] [3] with the OHSUMED document collection [12]. The documents have a similar nature to our data set [section 6]. Table 4 summaries the predicative accuracy from other researchers. The algorithms, data sets, and classificative problems are of course not strictly comparable. Nevertheless they do give some sense of the relative performance of our algorithm.

Table 4. Experimental results from other papers

Corpus	KNN	Rocc	WH	WORD	FWA	VWA
HD big*	.56	.46	.59	.44	-	-
MeSH 23**	-	.398	.296	-	.526	.526

 HD big is a copy of OHSUMED with a subset of categories in the heart disease sub-domain (49 unique categories) used in [2]. 183,229 training documents are used for the experiment.

** MeSH 23 is a copy of OHSUMED with a subset of the 23 MeSH disease categories used in [3]. 12,000 training documents are used for the experiment.

This performance is particularly significant given that our training document set is obviously an incomplete training set, only covering a small proportion of the important terms and their distribution for these pre-defined categories, but the classifier still performs satisfactorily. A statistics-based term weight algorithm would be expected to perform particularly poorly in these circumstances. Thus our method has shown an ability to reduce the size of training document set for creating a classifier. In effect, through searching in the concept hierarchy, we have automatically discovered the concept level, that is, the taxonomic standards to assign training documents to pre-defined categories.

In representing objects through their principal componenets, principal component analysis (PCA) has similar aims to ours, but with important differences. One of the main disadvantages of PCA is that it is difficult to know how many principal components to keep, although some rules of thumb are applied in practice. Another issue is that the performance of PCA strongly depends on the quality of training data set. Guided by the domain concept hierarchy, our method provides a semantic approach to overcome or relieve these disadvantages.

Acknowledgments

The authors thank the U.S. National Library of Medicine for having provided the UMLS knowledge resources.

References

[1] G. Salton. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, 1989.

[2] Y. Yang. An Evaluation of Statistical Approaches to Text Categorization. Journal of Information Retrieval, Vol 1, No. 1/2, pp 67--88, 1999.

[3] S. Shankar and G. Karypis. Weight adjustment schemes for a centroid based classifier. Computer Science Technical Report TR00-035, Department of Computer Science, University of Minnesota, Minneapolis, Minnesota, 2000.

[4] K. Beyer, J. Goldstein, R. Ramakrishnan and U. Shaft. When is 'nearest neighbour' meaningful? In Proc. of ICDT-1999, Jerusalem, Israel, 1999, pages 217-235, 1999.

[5] A. Hinneburg, C.C. Aggarwal and D.A. Keim. "What is the nearest neighbour in high dimensional spaces?" In Proc. of the International Conference on Very Large Databases (VLDB), pages 506--515, Cairo, Egypt, Sept. 2000. Morgan Kaufmann.

[6] T. Gruber. A translation approach to portable ontology specifications. Knowledge Acquisition, 1993. 5:p. 199-220.

[7] http://www.nlm.nih.gov/research/umls/

[8] http://www.ncbi.nlm.nih.gov/pubmed/

[9] http://www.nlm.nih.gov/mesh/meshhome.html

[10] W. Hersh, C. Buckley, T.J. Leone, and D. Hickam. OHSUMED: An interactive retrieval evaluation and new large test collection for research. In SIGIR-94, pages 192-201, 1994.