Neural Network-based Document Clustering Using WordNet Ontologies

Chihli Hung[1] **and Stefan Wermter**[2]

[1]De Lin Institute of Technology, Taiwan
chihli@mail.educities.edu.tw

[2]Centre for Hybrid Intelligent Systems
School of Computing and Technology
University of Sunderland, UK
stefan.wermter@sunderland.ac.uk, www.sunderland.ac.uk

Abstract. Three novel text vector representation approaches for neural network based document clustering are proposed. The first is the extended significance vector model (ESVM), the second is the hypernym significance vector model (HSVM) and the last is the hybrid vector space model (HyM). ESVM extracts the relationship between words and their preferred classified labels. HSVM exploits a semantic relationship from the WordNet ontology. A more general term, the hypernym, substitutes for terms with similar concepts. This hypernym semantic relationship supplements the neural model in document clustering. HyM is a combination of a TFxIDF vector and a hypernym significance vector, which combines the advantages and reduces the disadvantages from both unsupervised and supervised vector representation approaches. According to our experiments, the self-organising map (SOM) model based on the HyM text vector representation approach is able to improve classification accuracy and to reduce the average quantization error (AQE) on 10,000 full-text articles.

Keywords: document clustering, neural news classification, WordNet and self-organising map (SOM)

1. Introduction

Nowadays the problem is often not to access text information but to select the relevant documents. By grouping similar sets of information, an organised document structure can reduce the search space and help users to access a number of related documents. Document organisation can be carried out by document classification and document clustering [Hearst, 1999]. Document classification is text processing that classifies a document into one or more pre-defined classes and is treated as an expectation of information organisation by users. Document clustering is text processing that groups documents according to the similarity measure of their pre-defined features.

If the results of document clustering are compared with a classification label, the accuracy depends on the difference between implicit factors of the classification label and explicit definitions of cluster features and similarities. That is, when the definition of the features of documents has been determined, the clustering technique identifies clusters of documents based on some explicit formal evaluation, i.e. the definition of similarity of documents. However, documents are not only classified on the basis of their feature representation but also on the basis of implicitly subjective human concepts. Therefore, purely unsupervised document clustering methods are sometimes unable to represent document classification labels hidden in the document corpus. For example, the accuracy of document clustering cannot be expected to be very good if documents that are pre-classified as different classes share many of the same features, i.e. words [Aggarwal et al., 1999]. On the other hand, two documents that really belong together may be pre-classified as different classes.

Figure 1 is an example which illustrates different decisions by document clustering and classification. Documents are represented as circles with numbers, and circles with the same filled colour are pre-classified as the same class. There are nine documents which are pre-classified as two classes: black circles and white circles. However, based on the similarities of document vectors - when the similarities are evaluated by the Euclidean distance – the nine documents form the two clusters in Figure 1a. The distance from document 1 to document 2 is smaller than that to document 5, so document 1 is clustered in the same cluster as document 2 [Figure 1b]. Without embedding any external knowledge into the clustering approach, it is hard for document 1 to be grouped with document 5 [Figure 1c].

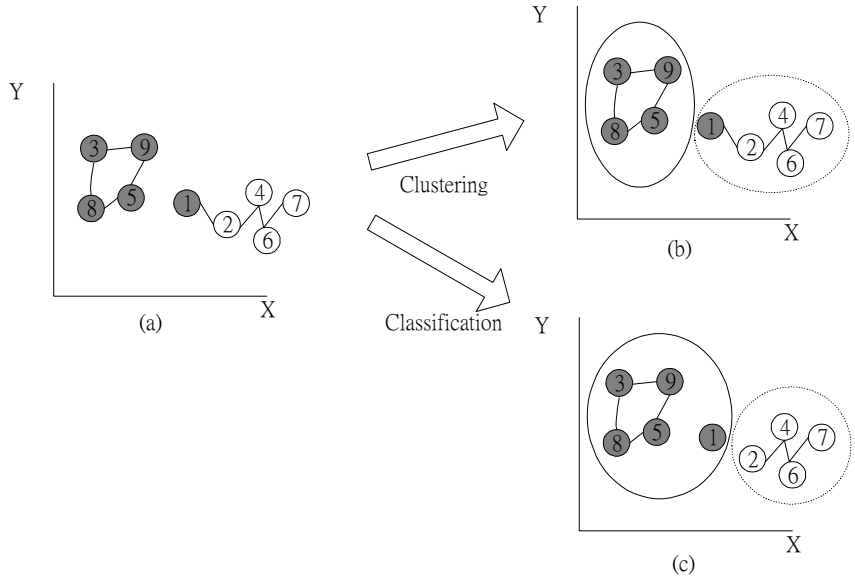


Figure 1. An example of different decisions from document clustering and human classification.

Therefore, the goal of this research is to build a hybrid model which applies the supervised and the unsupervised learning approaches to reduce the gap of document clustering and classification. Based on this hybrid model, we examine whether an unsupervised Self-Organising Map (SOM) [Kohonen, 1984] clustering model can be enhanced when document features, i.e., words, are guided by the relationship between words per se, their preferred classification labels and by the online lexical reference knowledge from the WordNet ontology [Miller, 1985].

The paper is structured as follows: There are several methods for a neural clustering model to integrate classification labels, which are introduced in Section 2. Sections 3-5 describe the three new vector representation approaches used by the hybrid SOM model to reduce the gap of inconsistency between supervised document classification and unsupervised document clustering. The extended significance vector model (ESVM) extracts the relationship of a word and its preferred semantic class in order to improve an external quantitative cluster criterion, i.e. classification accuracy. The hypernym significance vector model (HSVM) extracts semantic knowledge from an online lexical reference knowledge base, such as WordNet, to further discriminate word concepts from classes. The hybrid vector space model (HyM) uses the principle of supervised SOM learning [Honkela et al., 1996] and combines unsupervised and supervised vector representations, in order to eliminate the bias to show only the majority of classes and omit some minor classes of documents. Section 6 describes the distribution of the Reuters news corpus, which is used as the test bed in this research. Section 7 introduces the evaluation criteria used in this paper and we show our experimental results in Section 8.

2. Document Clustering Using Class Knowledge

Classification is supervised categorisation when classes are known, while clustering is unsupervised categorisation when classes are not known. Document clustering does not necessarily correspond to an existing classification since this technique relies on the representation of input features alone, e.g. words. However, the same document may be classified differently depending on a different purpose. For example, a news article which discusses the job losses of an international business in a country can be classified as economic news, social news, international business and labour news. In contrast to classification, document clustering groups similar news articles together because they contain similar input features. One news article is transformed to one document vector and can only be clustered to one cluster if the clustering algorithm, document features and similarity measure have been determined. Thus, a possible solution to reduce the gap between the classification concept and data-driven clustering is to use domain knowledge to allow integration of supervised human classification subjectivity [Jain et al., 1999].

Kohonen et al. [2000] point out that “Obviously, one should provide the different words with such weights that reflect their significance or power of discrimination between the topics. If, however, the

documents have some topic classification which contains relevant information, the words can also be weighted according to their Shannon entropy over the set of document classes." A modified vector space model (VSM) which includes class information has been used in their WebSOM project [Honkela et al., 1996; Kohonen et al., 2000].

Both clustering and classification may benefit from the integration of prior external class knowledge, which reflects specific classification concepts or organisation goals [Kim and Lee, 2000]. The incorporation of domain knowledge into clustering can be applied in several phases. Figure 2 illustrates three main stages at which classification knowledge can be added.

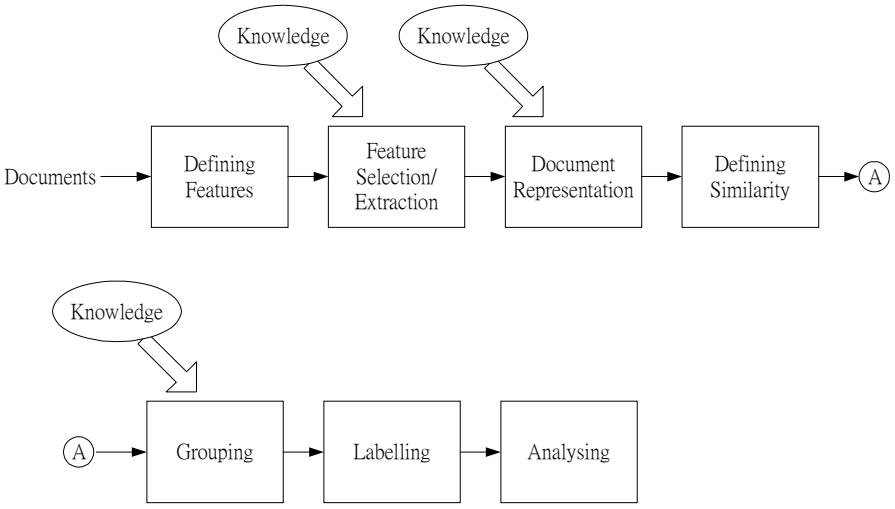


Figure 2. The stages of document clustering using human classification knowledge.

Applying classification knowledge at the stage of feature selection removes words with little discriminatory power between classes [Figure 2]. The general concept is that words which appear in many classes contain little discriminatory power between classes and vice versa. Thus, as an extreme example, words shown in only one class have the strongest discriminatory power. However, such words may be incorrectly spelt or too specific so that they are not general enough for the models. Therefore, approaches in this group usually apply to words that are not distributed evenly across all classes.

For example, Goldberg [1995] applies cue validity, Weiss et al. [1996] use term selection and Aggarwal et al. [1999] use the gini index of the word to select words with sufficient discriminatory ability. The advantages of these methods are a reduction of document vector dimensionality and an improvement of classification or clustering. However, to decide about the proper thresholds, i.e. a minimum threshold and a maximum threshold for choosing words which do have a relationship with the class, is not an easy task.

Another approach is to embed classification knowledge into the vector representation [Figure 2]. Vectors containing domain knowledge may improve the accuracy for an automated classifier and a clustering model. The general idea is to extract classification knowledge to produce vectors with some specific characteristic for a class. In other words, different words with a similar relationship to the same pre-classified class should have some similar characteristics in their vector representations.

For example, Wermter et al. [Wermter, 2000; Garfield and Wermter, 2002] define a semantic vector representation to embed class knowledge into the vector representation for an automated phrase and sentence classifier. Furthermore, Kohonen [2001] proposes the supervised SOM principle by concatenating a document vector with an extra class vector. The well-known WebSOM project [Honkela et al., 1996; Kohonen et al., 2000] applies pre-classified class information to a document map by using the entropy weighting method.

However, the semantic vector is a word vector and is not directly applicable to a document vector. A document vector using the Self-Organising Semantic Map [Ritter and Kohonen, 1989] or the supervised SOM principle [Honkela et al., 1996] needs to decide a proper weight value for the class part of the vector. A very small weight value does not offer the document vector enough class discriminatory ability, while a very large weight value loses the effect of the other part of the document vector [Bezdek and Pal, 1995]. Furthermore, the WebSOM entropy is usually used with the average random weighting approach whose

length of word vector is dependent on trial-and-error [Ritter and Kohonen, 1989; Hodge and Austin, 2002].

Classification knowledge can be used in the clustering algorithm, which groups documents by the influence of domain knowledge. For example, Arous and Ellouze [2003] develop a combined clustering approach by fusing the supervised SOM with other unsupervised SOMs, which produces the classification decision by majority voting. Kohonen [2001] proposes the learning vector quantization (LVQ) model which treats classification knowledge as guidance for the update rule to organise the output units of the SOM-like model. Mavroudi et al. [2003] propose the Supervised Network Self Organising Map (SNet-SOM) by using a hybrid unsupervised and supervised cost function. Although these models improve the classification accuracy, they lose some of the benefits of the characteristics of self-organising neural models, which represent similar documents by the same unit and deploy similar units in its neighbourhood.

3. Extracting Knowledge from Relationships between Words and their Preferred Classification Labels

Since the dimensionality for a document vector based on the vector space model (VSM) representation is the total number of different words in the target document set, the traditional VSM suffers from the massive size of the data set. In order to address this problem and to integrate the classification knowledge and data-driven clustering, the extended significance vector model (ESVM), which extends previous work on significance vectors [Wermter, 1995], is proposed in this section. Like the semantic vector representation approach [Wermter, 2000; Garfield and Wermter, 2002], the ESVM represents a preference for a specific semantic class, which is based on the assumption that the higher the relative frequency of the word for a class, the stronger the relationship between this word and its associated class. Unlike the semantic vector representation approach, the ESVM is designed for document vectors instead of word vectors [Wermter and Hung, 2002]. ESVM starts with the word-class occurrence matrix which can be depicted as:

$$\text{words} \downarrow \quad \text{classes} \rightarrow \quad \begin{bmatrix} o_{11} & o_{12} & o_{13}.....o_{1C} \\ o_{21} & o_{22} & o_{23}.....o_{2C} \\ ..... \\ ..... \\ o_{M1} & o_{M2} & o_{M3}.....o_{MC} \end{bmatrix},$$

where o_{ij} is the occurrence of word i shown in class j, C is the total number of classes and M is the total number of different words. An element of a significance word vector for a word i in class j is represented as w_{ij} and is obtained using Equation 1.

$$w_{ij} = \frac{o_{ij}}{\sum_{c=1}^{C} o_{ic}}, \tag{1}$$

where C is the total number of classes and $c \in [1..C]$.

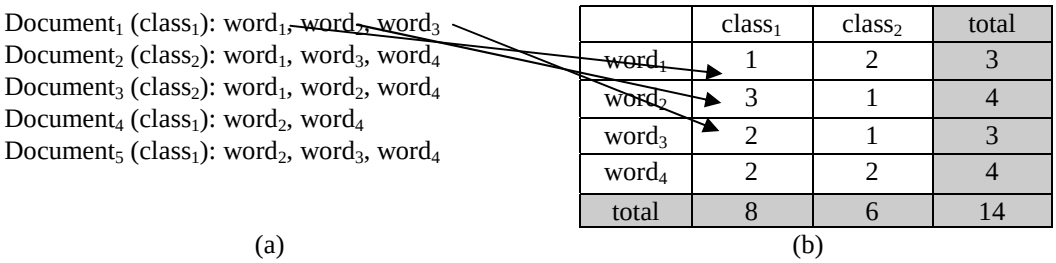
Equation 1 can be influenced by the different number of news documents observed in each class. When a specific class j contains significantly more articles than others, a word i may contain significantly more occurrences in class j than in other classes. Therefore, words may have the same significant class j and lose the discriminatory power between classes. Equation 2 is defined as the extended significance vector, which uses the logarithmic weights of the total number of word occurrences in the data set divided by the total number of word occurrences in a specific semantic class to alleviate skewed distributions in Equation 1. A more prominent class which contains more word occurrences will have smaller logarithmic values. Thus, the definition of an element in word vector $\vec{w}$ for class j is:

$$w_{ij} = \frac{o_{ij}}{\sum_{c=1}^{C} o_{ic}} \times \log \frac{\sum_{m=1}^{M}\sum_{c=1}^{C} o_{mc}}{\sum_{m=1}^{M} o_{mj}}, \quad (2)$$

where C is the total number of classes, M is the total number of different words, $m \in [1..M]$ and $c \in [1..C]$.

The news document vector $\vec{x}$ is then defined as the summation of extended significance word vectors $\vec{w}_i = (o_{i1}\ o_{i2}\ .....o_{iC})$ divided by the number of words in a document, which is defined as Equation 3.

$$\vec{x} = \frac{1}{s}\sum \vec{w}, \text{ where } s \text{ is the number of words in news document } x. \quad (3)$$



(a)

	class$_1$	class$_2$	total
word$_1$	1	2	3
word$_2$	3	1	4
word$_3$	2	1	3
word$_4$	2	2	4
total	8	6	14

(b)

	class$_1$	class$_2$
word$_1$	$\frac{1}{3}$	$\frac{2}{3}$
word$_2$	$\frac{3}{4}$	$\frac{1}{4}$
word$_3$	$\frac{2}{3}$	$\frac{1}{3}$
word$_4$	$\frac{2}{4}$	$\frac{2}{4}$

(c)

	class$_1$	class$_2$
word$_1$	$\frac{1}{3}\times\log\frac{14}{8}$	$\frac{2}{3}\times\log\frac{14}{6}$
word$_2$	$\frac{3}{4}\times\log\frac{14}{8}$	$\frac{1}{4}\times\log\frac{14}{6}$
word$_3$	$\frac{2}{3}\times\log\frac{14}{8}$	$\frac{1}{3}\times\log\frac{14}{6}$
word$_4$	$\frac{2}{4}\times\log\frac{14}{8}$	$\frac{2}{4}\times\log\frac{14}{6}$

(d)

	class$_1$	class$_2$
Document1	0.9793	1.0591
Document2	0.8394	1.2709
Document3	0.8861	1.2003
Document4	0.6995	0.6355
Document5	1.0730	0.9179

(e)

	class$_1$	class$_2$
Document1	0.3264	0.3530
Document2	0.2798	0.4236
Document3	0.2954	0.4001
Document4	0.3498	0.3177
Document5	0.3575	0.3060

(f)

Figure 3. An example of the extended significance vector model (ESVM) representation.

Using ESVM, each news article is represented by a C-dimensional vector, where C is the number of pre-classified classes instead of the number of different words in the master word list. This method overcomes the dimensionality problem by making use of a domain-dependent but automatically generated lexicon, since the total number of different classes C is usually much smaller than the number of different

words *M*. An example of the extended significance vector model (ESVM) representation is shown in Figure 3. In this example, there are five pre-classified documents and two classes [Figure 3a]. The pre-classified class for each document is shown in brackets. The word-class occurrence matrix is built by five documents [Figure 3b]. In Figure 3c, the significance word vector is built according to Equation 1. In Figure 3d, the extended significance word vector is built according to Equation 2. The document vector is composed by summing up the corresponding extended significance word vectors [Figure 3e]. Finally, the ESVM document vector is built by dividing by the number of different words shown in each document (Equation 3).

4. Extracting Knowledge from WordNet Ontologies

Theoretically, a cluster is a more general representation concept for its members and cluster members contain more specific concepts for their associated cluster. Based on this concept, an online lexical reference knowledge base which defines a hierarchically semantic relationship among words may help to improve the performance of clustering or classification. In other words, if a word is represented by another word which contains more general concepts in the same category, the concept of the document with replaced words will be more general and more similar to its cluster.

For instance, there are two groups of products in the example of a semantic word hierarchy in Figure 4. The first one is an edible fruit class which consists of citrus and apple fruits and the second one is a vegetable class which contains potato and carrot. Without using the semantic word hierarchy, to cluster or classify these four products requires the definition of some features with some discriminatory power, for example colour, shape etc. However, citrus and apple are kinds of edible fruit and potato and carrot are kinds of root vegetable. If the edible fruit substitutes for citrus and apple and the root vegetable substitutes for potato and carrot, there is no need to discriminate between citrus and apple and between potato and carrot. Thus, the original task becomes to cluster or classify two products, and this is much easier.

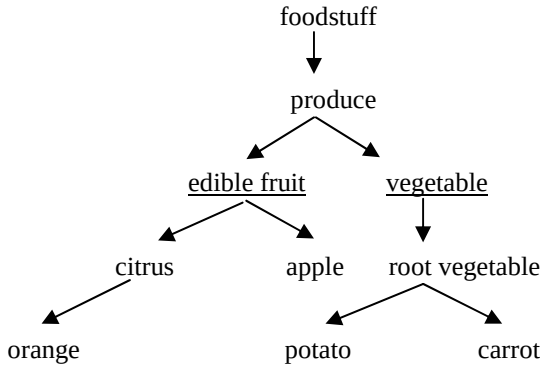


Figure 4. An example of hierarchically semantic relationships among words.

According to the concept of using semantic knowledge, we propose the hypernym significance vector model (HSVM). HSVM is based on the significance vector representation approach and further extracts the hypernym-hyponym relationship from the WordNet ontology [Miller, 1985]. WordNet is a well-known online lexical reference knowledge base and contains the semantic relationships from synset, a set of synonyms representing a distinct concept. WordNet (version 1.6[1]) contains 99,642 terms and 173,941 synsets, which are divided into four open-class categories (66,025 nouns, 12,127 verbs, 17,915 adjectives and 3,575 adverbs).

A hypernym of a term is a more general term and a hyponym is a more specific term. For example, in Figure 4, an apple is a hyponym of edible fruit and an edible fruit is a hypernym of an apple. This hypernym relationship from WordNet is exploited to examine whether fewer but more general concepts which substitute for original concepts can improve the performance of the SOM model.

[1] This research uses WordNet 1.6. The latest version of WordNet is WordNet 2.0, which can be downloaded from http://www.cogsci.princeton.edu/~wn/wn2.0.shtml

A word in different contexts may contain different concepts, and thus a word may be placed in different synsets, which form different hypernym trees. For example, to look up the hypernym of the word orange with the colour concept, the left hypernym tree in Figure 5 should be followed; otherwise the right hypernym tree should be followed. The 2-level hypernym for orange with the colour concept is color but with the fruit concept it is edible fruit.

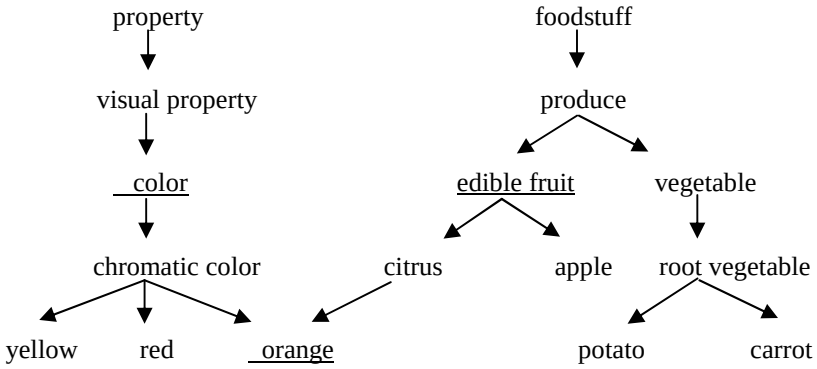


Figure 5. An example of two hypernym trees for the term orange.

It is hard to determine the right concept for an ambiguous word from several synsets and it is hard to decide the concept of a document that contains several ambiguous terms. Brezeale [1999] directly uses the first synset from WordNet because of the greatest frequency of occurrence in WordNet. Voorhees [1993] proposes a method called hood to resolve this difficulty. An ambiguous word looks for some level of hypernym until it finds the same hypernym in each hypernym tree. A hood is defined as the direct descendent of this same hypernym which is shared by different concepts of a term. The meaning of ambiguous words can be decided by counting the number of other words in the text that occur in each of the different sense's hoods. Then the specific hood with the largest number is represented as the sense of the ambiguous word. Scott and Matwin [1998] used hypernym density to decide which synset is more likely than others to represent the document. The hypernym density is defined as the number of occurrences of the synset within the document divided by the number of words in the document. The synset with the higher density value is more suitable to represent the document.

This study does not use the synset directly but takes advantage of the synset's gloss, because two synonyms may not co-occur in a document, for example, color and colour, and orange and orangeness. The synset's gloss contains an explanation of the meaning and an example sentence of each concept and can be treated as a small piece of the document with a core meaning. For example, the gloss of the word "orange", with the fruit concept is "round yellow to orange fruit of any of several citrus trees" and with the colour concept is "any of a range of colors between red and yellow". In contrast to synonyms, words in the gloss and their target word may be more likely to co-occur.

We want to use the hypernym relationship from WordNet to enhance our significance vector model even further. The hypernym significance vector model (HSVM) is formed using the following steps. First, the semantic lexicon based on ESVM is converted into its n-level hypernym version, which is carried out by looking up the n-level hypernym for each word in each class. Each ambiguous word in the original lexicon contains several senses and each sense has its own gloss. Each gloss of a word is transformed into a vector using the extended significance vector representation approach and this transformation is performed based on the class, which means that other classes are ignored when transforming words in a specific class.

Second, to decide the possible gloss for an ambiguous word, the specific element value of each gloss vector in the specific class of the original semantic lexicon is compared. The gloss vector with the highest value in the specific element to represent the original word is chosen. For example, a comparison is made for the first element only when transforming words in class 1. The second element is compared when transforming words in class 2 and so on.

To illustrate this approach, we give the following example. Assume that two gloss significance vectors for the word orange with colour concept and with fruit concept are [0.101 0.203 0.302 … 0.031] and [0.201 0.103 0.222 … 0.021] respectively. When orange in class 1 is converted into its hypernym, only the first element is compared for two gloss vectors. Thus, the gloss with fruit concept is chosen for orange in class 1 since the first element in the gloss vector with fruit concept is greater than that with colour concept

(0.201>0.101). When orange in class 2 is converted into its hypernym, the colour concept is chosen (0.203>0.103). One word in one class has only one hypernym tree and one word in different classes may share the same hypernym tree.

This procedure is different from word sense disambiguation (WSD), which usually considers word meaning from contexts. In our approach, a word occurring in documents that are pre-classified as the same class has the same representation. This approach helps us to investigate the relationship between words and their associated classes to reduce the gap of inconsistent decisions from automated clustering and human classification.

Third, going up n-levels in the hypernym tree, this hypernym is used to build the hypernym version of a semantic lexicon for all words in all classes, and a word-hypernym look-up table is also built. Fourth, each news article is converted from its original version into its n-level hypernym version. Since a look-up table of word-hypernyms has been built at the previous stage, each word in each news document is converted based on its pre-classification class. Finally, the n-level hypernym data set is transformed to vectors by using the extended significance vector representation.

5. The Hybrid Text Vector Representation

SOM-like learning tends to represent the number of input vectors by output units whose number is roughly proportional to the number of input vectors [Kohonen, 2001]. Thus, minor classes may be overwhelmed by major classes when they are labelled by class identities and are mapped to the same unit. For example, in the WebSOM project [Kohonen et al., 2000], the units of a SOM representing Usenet newsgroup articles are labelled with the name of the newsgroup of the majority of articles mapped onto this specific unit. This is a potential weakness when applying a SOM-like model as an interface for searching a collection of documents, because several minor class labels cannot be shown on the map, but these documents may be search targets for users. A SOM-like model which represents the class labels, such as class number, class terms etc., suffers from this problem when searching an uneven distribution of documents in classes. Thus, the SOM model based on ESVM or HSVM cannot explain the concept of an output map using significant terms since their elements still illustrate significant classes or pre-classified classes, which may not be detailed enough.

An alternative is to consider a SOM-like map as a semantic geographical map that reflects the significance of terms in different areas which are shared by some major and minor classes. A term represents documents usually because this term is important for those documents. The important term is usually transformed to a greater element value based on the vector space model (VSM), such as term frequency (TF) or term frequency x inverse document frequency (TFxIDF). Therefore, the element with a greater value in a unit vector can be used as a representative term for documents that are mapped to this unit.

Roussinov and Chen [1999] assign a term to each output unit of the SOM map by choosing the unit element that contains the largest value. This method has been used by several researchers [Lin et al., 1991; Lin, 1997; Chen et al., 1996; Ritter and Kohonen, 1989] and is also used in this paper. Two terms whose weights are the most significant are used to represent the labels of the unit. Theoretically, neighbouring units in a SOM map represent similar concepts. Thus, some of the most significant terms in two neighbouring units should contain the same or similar concepts.

Inspired by the supervised SOM principle [Honkela et al., 1996], a document vector is formed by concatenating an unsupervised vector based on the vector space model (VSM) and a supervised vector based on the hypernym significance vector model (HSVM). This approach is called the hybrid vector space model (HyM), and intends to harmonise the advantages and eliminate the disadvantages from both unsupervised and supervised vector representation approaches. This is because the SOM model using HyM can take advantage of the unsupervised vector representation, i.e. VSM, which is able to represent the document collection based on a feature map, and also has the advantage of higher accuracy based on the supervised vector representation approach, i.e. the hypernym significance vector model (HSVM).

HyM is based on VSM and uses a pre-defined parameter, i.e. γ, to control the influence of a hypernym significance vector (Equation 4). When the control parameter γ is larger, the effect of the supervised part of the document vector is more significant. In the extreme example, when the value of γ is 1, this is a hypernym significance vector representation approach and when the value of γ is zero, this is a traditional vector space representation approach.

$$HyM = \begin{bmatrix} (1-\gamma)\cdot VSM \\ \gamma \cdot HSVM \end{bmatrix} = \begin{bmatrix} (1-\gamma)\cdot VSM \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ \gamma \cdot HSVM \end{bmatrix}, \quad (4)$$

where γ is between 0 and 1.

6. Reuters Corpus of News Articles

We test our approaches by using the current version of the Reuters news corpus, RCV1[2], since this series of news corpora is a representative test for text classification, a common benchmark and a recent comprehensive data source [Sebastiani, 2002]. This corpus is made up of 806,791 news articles from issues of Reuters between the 20th August, 1996 and 19th August, 1997 and contains about two hundred million word occurrences and ten million paragraphs.

Each document is saved in a standard XML format and is pre-classified by three different codes of categories, which are industry code, region code and topic code. There are 870, 366 and 126 categories for the industry, region and topic code respectively. Only the topic categories are used in this research. Among 126 topic categories, 23 topic categories contain no news articles. A news article can be pre-classified as more than one topic or no topic. In this corpus, each news article is pre-classified as 3.17 topics on average.

This research concentrates on the eight most prominent topics [Table 1] because the previous version of Reuters Corpus, i.e. Reuters-21578, is also divided into eight main topics which contain 135 detailed topics. Even though Reuters-RCV1 has not become a research standard yet, it is believed that this Reuters news corpus will be used as a benchmark in the near future.

Table 1. The description of chosen topics and their distribution over the whole Reuters-RCV1 corpus.

Topic	Description	Distribution	Percentage
C15	Performance	149,359	11.44%
C151	Accounts/Earnings	81,201	6.22%
C152	Comment/Forecasts	72,910	5.58%
CCAT	Corporate/Industrial	372,099	28.50%
ECAT	Economics	116,207	8.90%
GCAT	Government/Social	232,032	17.77%
M14	Commodity markets	84,085	6.44%
MCAT	Markets	197,813	15.15%
Total		1,305,706	100.00%

Since a news article can be pre-classified as more than one topic, the multi-topic composition is treated as a new topic and as a new class in this research. Thus the 8 chosen topics are expanded into 40 combined topics for the first 10,000 news articles [Table 2]. For example, topic composition 1 or class 1 is a combination of topics ECAT and MCAT. News articles that are pre-classified as this class are articles about economics and markets.

[2] The Reuters-RCV1 corpus can be found at http://about.reuters.com/researchandstandards/corpus/

Table 2. The distribution of classes for 10,000 full-text news data set.

The 10,000 full-text news data set					
No	Class	#	No	Class	#
1	ECAT/MCAT	155	21	C15/C152/CCAT/MCAT	20
2	CCAT	1,780	22	C15/C151/CCAT/ECAT	2
3	C15/C151/CCAT/ECAT/GCAT	6	23	CCAT/ECAT/MCAT	7
4	C15/C151/CCAT	999	24	C15/C152/CCAT/ECAT/M14/MCAT	1
5	M14/MCAT	877	25	C15/CCAT	1
6	ECAT	771	26	CCAT/GCAT/M14/MCAT	31
7	CCAT/GCAT	293	27	GCAT/M14/MCAT	4
8	CCAT/ECAT/GCAT	162	28	C15/C152/CCAT/ECAT/GCAT	2
9	MCAT	1,135	29	ECAT/GCAT/MCAT	17
10	GCAT	2,152	30	C15/C152/CCAT/M14/MCAT	2
11	ECAT/GCAT	195	31	C15/C152/CCAT/ECAT	3
12	C15/C152/CCAT	802	32	C151/C152	1
13	CCAT/ECAT	157	33	CCAT/ECAT/GCAT/MCAT	1
14	C151	76	34	GCAT/MCAT	6
15	C152	25	35	CCAT/ECAT/M14/MCAT	6
16	CCAT/M14/MCAT	210	36	CCAT/ECAT/GCAT/M14/MCAT	4
17	C15/C152/CCAT/GCAT	9	37	C15/C151/CCAT/MCAT	1
18	CCAT/MCAT	28	38	M14	1
19	C15/C151/C152/CCAT	52	39	C15/C151/CCAT/GCAT	1
20	ECAT/M14/MCAT	4	40	C15/C152/CCAT/ECAT/MCAT	1

7. Evaluation Criteria

The evaluation of SOM-like models needs some careful analysis. The unsupervised feature of SOMs usually requires the inclusion of the subjective judgements of domain experts [Roussinov and Chen, 1999]. Even though it is possible to see clusters through the SOM-like maps, human qualitative judgements should not be the only evaluation criterion. The main reason is that human judgements are so subjective that different assessments may be made by the same person at a different time or for a different process.

Unlike qualitative assessment, quantitative criteria can be divided into two types: internal and external [Steinbach et al., 2000]. The internal quantitative measure is data-driven and the quantization error is applied in this research. The external quantitative measure evaluates how well the clustering model matches some prior knowledge which is usually provided by humans. The most common form of such external information is human manual classification knowledge, so classification accuracy is used in this research. These two evaluation criteria have been used by several researchers in the field of SOM clustering [Kohonen et al., 2000; Choudhary and Bhattacharyya, 2002] and are also used in our research.

7.1 Quantization Error

The quantization error (QE) is suggested by Kohonen as a measurement used in the vector quantization technique [Kohonen, 2001]. The QE, also called the distortion measure, is defined as the sum of the Euclidean distance between every input vector and its best matching unit (BMU). Given a data set X containing input vectors x_i, the QE is described as Equation 5.

$$QE = \sum_{i=1}^{N} \left\| x_i - w_i \right\|, \tag{5}$$

where w_i is the weight vector of BMU for input sample i and N is the total number of input vectors.

Small values of the QE cause small distortion for all input vectors to their cluster centres. All good clustering approaches should have a small QE, which is a direct index linked to the model's explanation

ability. There are several similar criteria, such as the mean quantization error (MQE) and the average quantization error (AQE). The MQE is a mean value of QE to the number of units (Equation 6). A bigger sized map has a smaller value of MQE where the two maps contain the same QEs. This criterion seems less useful for comparing different sizes of maps. The AQE (Equation 7) is the average value of QE to the number of input vectors rather than the number of units. It is an indicator of the quality of the model or the unit, which represents an average input vector, and is therefore used in this research.

$$MQE = \frac{QE}{U}, \tag{6}$$

where U is the number of units.

$$AQE = \frac{QE}{N}, \tag{7}$$

where N is the number of input vectors associated to a model or unit.

7.2 Classification Accuracy

A clustering model can also act as an unsupervised classifier [Aggarwal et al., 1999]. Like supervised classification, this criterion needs human involvement, either before or after clustering. Several researchers evaluate the performance of document clustering based on recall and precision, F-measure or classification accuracy [van Rijsbergen, 1979]. For example, Roussinov and Chen [1999] assess document clustering based on recall and precision. Wong et al. [2000] and Massey [2003] evaluate document clustering by the F-measure. Sahami et al. [1998], Kohonen et al. [2000], and Choudhary and Bhattacharyya [2002] evaluate text clustering by classification accuracy.

Kohonen et al. [2000] define the classification error thus: "all documents that represented a minority newsgroup at any grid point were counted as classification errors … the node and the abstracts belonging to the other subsections were considered as misclassifications." That is, each document has a pre-defined newsgroup label. After the training process, the category of a map unit is assigned according to the highest number of pre-defined labels of documents. Therefore, every unit represents its major article labels. The pre-defined label of each document which is mapped into this unit will be replaced by the unit label. Thus, if the unit label of each document matches its pre-defined label, it is a correct mapping. The classification accuracy is calculated from the number of correct mappings relative to the number of input articles.

For example, consider 10 news articles in the data set and 1 unit in a trained SOM. Three articles are pre-classified as class 1 and seven articles are pre-classified as class 2. All news articles are mapped to unit 1 because there is only one unit in this example SOM. Thus, the class 2 label is assigned to unit 1 and all news articles which are mapped to unit 1 are assigned as the unit label class 2. In this case, classification accuracy is 70% since three out of ten articles are assigned to different labels from their pre-classified labels. If five news articles are pre-classified to topic 1 and the rest of the articles are pre-classified as topic 2, classification accuracy is 50%.

A document sometimes can be pre-classified as more than one topic. The documents in the Reuters news collection are one example. Several researchers avoid the multi-topic problem by choosing documents that are pre-classified as only one topic for the text clustering or classification task. For example, Sahami et al. [1998] select documents that are pre-classified as only one topic from a subset of the pre-defined topics from the Reuters-22173. Kim and Lee [2000] use a controlled subset of the Reuters-21578 to cluster documents that have a single pre-classified topic in order to avoid the ambiguity of multiple topics. Arevian et al. [2003] use eight main topics instead of 135 detailed topics from the Reuters-21578, which reduces the possibilities of ambiguous multiple topics.

It is necessary to give a clear definition of the correct mapping for the multi-topic clustering task. One option is to match one to many pre-classified labels [Wermter and Hung, 2002; Massey, 2003]. In other words, no matter how many pre-classified labels a document has, a correct mapping is achieved if the topic label of each document belongs to the subset of its pre-classified labels. This definition is sometimes inexact because the possibility of a correct mapping for a combined distribution is greater than for a unique one. For example, a document which is pre-classified as CCAT and GCAT contains a probability of classification accuracy up to 46.27% (28.50%+17.77%) because the frequency of CCAT and GCAT are

28.50% and 17.77% respectively [see Table 1].

In this research, a stricter definition of the correct mapping is used. Any multi-topic combination which is assigned to a document is treated as a new topic or class. That is, all multi-topic combinations are replaced by new classes. As in the previous example, the CCAT and GCAT combined topic is assigned to class 7 whose frequency is merely 2.93% ($\frac{293}{10{,}000}$) [see Table 2].

8. Experiments

8.1 Preprocessing

We test our hybrid models on the Reuters-RCV1 corpus. Extracting full-text from raw data in XML format is the first phase, and then documents that are pre-assigned to one or several of the eight most prominent topics are included [Table 1]. The traditional vector space model (VSM) [Salton, 1989] is used to represent a full-text document and can be treated as a baseline since it is one of the best-known text vector representation approaches. However, this method is likely to suffer from the curse of dimensionality because the dimensionality of the document-word matrix is the total number of different words. Thus some feature selection techniques are useful when dealing with a large data set. A common strategy is used that removes those common words which are described in a stop word list from the WordNet package, lemmatises words to their base forms and restricts itself to words found in a machine readable dictionary, such as WordNet [Miller, 1985]. WordNet only includes open-classed words, i.e. nouns, verbs, adjectives and adverbs, as these words are believed to be able to convey enough information about document concepts. After this pre-processing, for the first 10,000 documents in the Reuters-RCV1, there are 16,122 distinct words in the master word list, which form a 10,000 x 16,122 document-word matrix. It still requires a lot of resources e.g. memory and CPU time, to handle a clustering task with a matrix of this size, so a further feature selection approach is useful. As in the work of Chen et al. [1996], only the 1,000 most frequent words from the master word list are used in this research since this method has provided the greatest overlap in representations [Roussinov and Chen, 1999] and has been shown to be as good as most dimensionality reduction techniques [Schütze and Silverstein, 1997; Chakrabarti, 2000].

8.2 Experiment Design

Due to the characteristics of SOM neural clustering models, some parameters need to be defined in advance. For example, the SOM requires a pre-defined architecture of a topographic map, training length, and learning rate. To decide on the SOM topographical structure, various structures between 10x10 and 25x25 have been considered. It is found that the performance of the 15x15 map is better than that of the 10x10 map and is comparable to that of the 25x25 map for 10,000 full-text documents in terms of classification accuracy and AQE. To decide on the SOM training length, different training lengths have also been tested and we found that after 50,000 iterations a longer training length does not produce a significant improvement for 10,000 full-text documents.

The learning rate and neighbouring size decay over time to pursue the convergence of the model [Mehrotra et al., 1997] and therefore the training length needs to be pre-defined as a stop criterion for the SOM. According to Kohonen's suggestions [2001], the learning rate should not start from a very large value and the initial neighbouring size should encompass the whole map. A very large initial learning rate may make the model unstable and a too small initial learning rate may only achieve local minima. In this research, an initial learning rate, i.e. 0.1, is chosen for the SOM model and decays to 0.001 over a pre-defined training time. This initial learning rate is chosen experimentally and the same initial learning rate is used for all models to remove the effect of different learning rates on models.

8.3 SOM Models using VSM and ESVM

The SOM model based on the normalised TFxIDF enforces unsupervised learning since this method does not include any external classification knowledge. The SOM model based on the extended significance vector model (ESVM), which applies human classification knowledge implicitly, can be treated as a guided

self-organising model. This model is based on the lexicon of significance vectors, and it is superior to those with the vector space model (VSM) [Table 3]. These results show that the pre-classified information offers better class discriminatory power for the SOM based on ESVM. Since the dimensionality of ESVM is only the number of classes, i.e. 40, which is much smaller than that of VSM, the constraint of using the 1,000 most frequent words can be relaxed. Thus, all 16,122 distinct words in the master word list are used for ESVM and the hypernym significance vector model (HSVM) in this research. According to Table 3, the SOM shows a greater average quantization error (AQE) when it is based on TFxIDF. The reason for this outcome is because the dimensionality of a document vector using the VSM representation is equal to the total number of words so more words produce larger quantization error. Therefore, a comparison of AQEs of models using different dimensionalities of vectors is not useful.

Table 3. A comparison of SOMs based on TFxIDF and ESVM evaluated by classification accuracy and AQE.

	TFxIDF	ESVM
Classification Accuracy	69.15%	81.28%
AQE	0.932	0.136

8.4 Different Levels of the WordNet Hypernym Tree

WordNet contains hypernym hierarchy trees for nouns and verbs. The higher the level of the hypernym, the more general the meaning of the concept. We have 16,122 different words for the significance vector representation approach but the total number of words is reduced to 8,683, 5,992, and 4,733 for using the 1-, 2- and 3-level hypernym significance vector representation approach respectively. If the level is too high, different senses of words may be treated as the same word and thus lose the discriminatory power between classes. This subsection investigates the performance of models using different levels of hypernyms. For convenience, the 1-, 2- and 3-level hypernym significance vector model is named HSVM1, HSVM2 and HSVM3, respectively.

According to the experimental results, the classification accuracy for the SOM model using the 1-, 2- and 3-level hypernym vector representation approach is 95.28%, 92.89% and 88.18% respectively. Therefore, the 1-level hypernym significance vector representation approach is used for a comparison of the SOM models based on VSM and ESVM. The results of the experiments show that the integration of a neural model and WordNet knowledge outperforms the neural model without WordNet knowledge.

8.5 SOM Models using HyM

According to our experiments in Subsections 8.3 and 8.4, the SOM model based on the 1-level hypernym significance vector representation approach achieves the highest accuracy. Therefore, the hybrid vector space model (HyM) integrates the TFxIDF vector representation approach with the 1-level hypernym significance vector representation approach in order to harmonise the advantages from both vector representations as mentioned in Section 5.

Since the purpose of this subsection is to examine the performance of the hybrid vector space model (HyM), different values of γ are used. For convenience, the SOM model using HyM with 0 for γ is termed HyM00, and the SOM model using HyM with 0.2 for γ is termed HyM02 and so on. Six different values of γ are used in this subsection, which are 0, 0.2, 0.4, 0.6, 0.8 and 1. The values of accuracy for these SOM models are between 69.67% and 94.76% while the values of the average quantization error (AQE) are between 0.137 and 0.931 [Table 4]. A higher γ produces higher classification accuracy and a lower AQE. Compared to the SOM model using the hypernym significance vector representation approach, the SOM model using the HyM approach achieves comparable classification accuracy when the value of γ is about 0.6.

Table 4. A comparison of SOMs based on different γ of HyM evaluated by classification accuracy and AQE.

	HyM00	HyM02	HyM04	HyM06	HyM08	HyM10
Classification Accuracy	69.67%	73.45%	86.45%	94.10%	94.39%	94.76%
AQE	0.931	0.748	0.567	0.391	0.226	0.137

Even though the SOM model using HyM loses some degree of classification accuracy, a feature map is produced to illustrate a document collection, which diverts the deficiency of a bias for the majority of classes for a SOM-like map when it is used as a browsing interface. We show a SOM map using HyM06 and label its units by the major class number [Table 2]. In Figure 6, many minor classes cannot be seen in the map, which means that those minor classes cannot be found on the map.

10	10	10	10	10	10	7	7	7	2	2	2	2	2	2
10	10	10	10	10		7	7		2	2	2	2	2	2
10	10	10	10	10	11	7	8	17	5	2	2	2	2	2
10	10	10	10	10	11	8	8	8	13	2	2	2	2	2
10	10	10	10	10	11	11	8	8	13	13	18	2	2	2
10	10	10	10	10	29	11	11	13	13	6	6	2	6	6
10		10	10	12	22	21	18	18	13	13	13	6	6	6
10	28	14	12	12	12	12	14	16	13	13	6	6	6	6
4	4	19	12	12	12	12	19	16	16	20	6	6	6	6
4	14	12	12	12	12	12		16	5	6	1	6	6	6
4	4	4	12	12	12	12		16	16	16	6	1	1	9
4	4	12	12	12	12		5	5	5		21	9	9	9
4	4	4		12	12		5	5	5	5	5	9	9	9
4	4	4	12	12		5	5	5	5	5	9	9	9	9
4	4	4	4			5	5	5	5	5	5	9	9	9

Figure 6. A map of the SOM using the HyM06 representation approach labelled by major class number.

police russian	police party	party minis*	party minis*	party minis*	party court	court tobac*	court tobac*	tobac* court	comp* say	comp* say	comp* say	comp* say	comp* share	comp* share
police russian	police party	party minis*	party minis*	party minis*		court tobac*	tobac* court		comp* say	comp* say	comp* million	comp* million	comp* share	share comp*
police russian	police minis*	minis* party	minis* party	minis* party	minis* gover*	gover* union	union gover*	comp* say	comp* say	comp* say	comp* million	comp* million	comp* share	share comp*
play win	win play	story minis*	story minis*	minis* gover*	gover* union	union budget	union budget	say comp*	comp* percent	comp* percent	percent comp*	million percent	million share	share bank
play match	play match	play win	story minis*	minis* story	gover* union	budget union	budget union	budget union	percent budget	percent rate	percent rate	percent million	percent bank	bond bank
play match	play match	play match	percent year	percent year	budget percent	budget union	budget percent	budget percent	rate percent	rate percent	rate percent	percent rate	bond percent	bond coupon
play match		profit result	profit share	share percent	share percent	share percent	percent budget	percent budget	rate percent	rate percent	rate percent	rate percent	bond percent	bond percent
profit result	profit million	profit share	share profit	share profit	share stock	share percent	percent tonne	toune percent	rate percent	rate percent	rate percent	rate percent	bond percent	bond coupon
profit million	profit million	profit million	profit share	share profit	share analyst	share stock	share tonne	tonne percent	tonne percent	rate tonne	rate percent	rate percent	rate percent	bond percent
profit million	profit million	profit share	share profit	share analyst	share stock	share stock		tonne wheat	tonne wheat	rate percent	rate percent	rate percent	rate bond	bond percent
profit net	profit million	profit million	share profit	share profit	share analyst	share stock		tonne wheat	tonne wheat	tonne wheat	rate percent	rate bank	rate bank	rate bank
profit net	profit net	profit yen	yen share	share yen	share yen		tonne wheat	tonne wheat	tonne wheat		rate bank	rate bank	rate bank	rate bank
net loss	net loss	yen net		yen share	yen share		tonne price	tonne wheat	bond trader	bond trader	rate market	rate bank	rate bank	rate bank
net loss	net loss	yen net	yen net	yen specify		tonne cent	tonne cent	tonne cent	trader tonne	trader market	rate market	rate bank	rate bank	rate bank
net loss	net loss	yen net	yen net			yen cent	cent tonne	cent trader	cent trader	trader cent	trader market	trader market	rate bank	rate bank

Figure 7. A map of the SOM using the HyM06 representation approach labelled by two significance terms. A word containing a star "*" is an abbreviation, for example, "minis*" for "minister", "tobac*" for "tobacco", "comp*" for "company", "gover*" for "government".

In contrast to this labelling approach, two terms out of 1,000 index words whose weights are the most significant in each unit are used to represent the labels of the unit in the SOM map [Figure 7]. Two neighbouring units in the SOM map are represented by one identical word or one related word, which demonstrates that units in a neighbourhood represent similar concepts and those concepts are altered smoothly. For example, units on the top left mainly discuss police and party, the units on the top right discuss company and share issues, the bottom right units are related to bank and rate situations and the bottom left units discuss performance matters.

9. Conclusion

The main purpose of this paper is to induce additional semantic category knowledge into the SOM model to enhance domain clustering performance. We propose three novel vector representation approaches, i.e. the extended significance vector model (ESVM), the hypernym significance vector model (HSVM) and the hybrid vector space model (HyM) in this paper. The SOM model based on ESVM extracts the relationship between words and their preferred classification labels. This approach is able to avoid the curse of dimensionality and achieve higher classification accuracy than the SOM based on the traditional vector space model (VSM). Based on the ESVM representation technique, we propose another novel vector representation approach, i.e. HSVM. The SOM model based on HSVM, extracting symbolic knowledge from the WordNet ontology, further improves the performance of clustering. Finally, we integrate the TFxIDF vector representation approach and the hypernym significance vector representation approach as a hybrid vector space model, which can be treated as an interface for document browsing. The SOM based on HyM offers a way to show the inner structure of a document collection by a semantic feature map, and also provides comparable classification accuracy to the SOM using the hypernym significance vector representation technique alone. These results demonstrate that knowledge from an ontology such as the WordNet ontology is able to enhance SOM clustering performance.

References

Aggarwal, C.C., Gates, S.C. and Yu, P.S. (1999), "On the merits of building categorization systems by supervised clustering," *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 352-356.

Arevian, G., Wermter, S. and Panchev, C. (2003), "Symbolic state transducers and recurrent neural preference machines for text mining," *International Journal on Approximate Reasoning*, vol. 32, no. 2/3, pp. 237-258.

Arous, N. and Ellouze, N. (2003), "Cooperative supervised and unsupervised learning algorithm for phoneme recognition in continuous speech and speaker-independent context, *Neurocomputing*, vol. 51, pp.225-235.

Bezdek, J.C. and Pal, N.R. (1995), "A note on self-organizing semantic maps," *IEEE Transactions on Neural Networks*, vol. 6, no. 5, September, pp. 1029-1036.

Brezeale, D. (1999), "The organization of Internet web pages using WordNet and self-organizing maps," *Masters thesis*, University of Texas at Arlington.

Chakrabarti, S. (2000), "Data mining for hypertext: a tutorial survey," *ACM SIGKDD Explorations*, vol. 1, no.2, pp. 1-11.

Chen, H., Schuffels, C. and Orwig, R. (1996), "Internet categorization and search: a self-organizing approach," *Journal of Visual Communication and Image Representation*, vol. 7, no. 1, pp. 88-102.

Choudhary, B. and Bhattacharyya, P. (2002), "Text clustering using semantics," *Proceedings of the 11th International World Wide Web Conference*, WWW2002, Honolulu, Hawaii, USA, http://www2002.org/CDROM/poster/79.pdf.

Garfield, S. and Wermter, S. (2002), "Recurrent neural learning for helpdesk call routing," *Proceedings of ICANN-2002, the International Conference on Artificial Neural Networks*, Madrid, Spain, August, pp. 296-301.

Goldberg, J.L. (1995), "CDM: an approach to learning in text categorization," *Proceedings of TAI 95 the 7th IEEE International Conference on Tools with Artificial Intelligence*, pp. 258-265.

Hearst, M. A. (1999), "The use of categories and clusters for organizing retrieval results," Strzalkowski, T. (ed.), *Natural Language Information Retrieval*, Kluwer Academic Publishers, Netherlands, pp.333-374.

Hodge, V.J. and Austin, J. (2002), "Hierarchical word clustering – automatic thesaurus generation.," *Neurocomputing*, vol. 48, pp. 819-864.

Honkela, T., Kaski, S., Lagus, K. and Kohonen, T. (1996), "Exploration of full-text databases with self-organizing maps," *Proceedings of the International Conference on Neural Networks (ICNN'96)*, Washington, pp.56-61.

Jain, A.K., Murty, M.N. and Flynn, P.J. (1999), "Data clustering: a review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264-323.

Kim, H-J. and Lee, S-G. (2000), "A semi-supervised document clustering technique for information organization," *Proceedings of the Ninth International Conference on Information and Knowledge Management*, pp. 30-37.

Kohonen, T. (1984), *Self-organization and associative memory*, Springer-Verlag, Berlin.

Kohonen, T. (2001), *Self-Organizing Maps*, 3rd edition. Springer-Verlag, Berline, Heidelberg, New York.

Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V. and Saarela, A. (2000), "Self organization of a massive document collection," *IEEE Transactions on Neural Networks*, vol. 11, no. 3, pp. 574-585.

Lin, X. (1997), "Map displays for information retrieval," *Journal of the American Society for Information Science*, vol. 58, no. 1, pp.40-54.

Lin, X., Soergel, D., Marchionini, G. (1991), "A Self-organizing Semantic Map for Information Retrieval," *Proceedings of the 14th Annual International ACM/SIGIR Conference on R & D in Information Retrieval*, pp. 262-269.

Massey, L. (2003), "Evaluating quality of text clustering with ART1," *Neural Networks*, vol. 16, no. 5-6, special issue: Advances in neural networks research -- IJCNN'03, pp. 771-778.

Mavroudi, S., Dragomir, A, Papadimitriou, S. and Bezerianos, A. (2003), "Integrating supervised and unsupervised learning in self organizing maps for gene expression data analysis," *ICANN/ICONIP 2003*, pp. 262-270.

Mehrotra, K., Mohan, C.K., and Ranka, S. (1997), *Elements of Artificial Neural Networks*, MIT press.

Miller, G.A. (1985), "WordNet: a dictionary browser," *Proceedings of the First International Conference on Information in Data*, University of Waterloo, Waterloo, pp. 25-28.

Van Rijsbergen, C.J. (1979), *Information Retrieval*, London, Butterworths, 2nd Edition.

Ritter, H. and Kohonen, T. (1989), "Self-organizing semantic maps," *Biol. Cybern.*, vol. 61, pp. 241-254.

Roussinov, D.G. and Chen, H. (1999), "Document clustering for electronic meetings: an experimental comparison of two techniques," *Decision Support Systems*, vol. 27, pp. 67-79.

Sahami, M., Yusufali, S. and Baldonado, Q.W. (1998), "SONIA: A service for organizing networked information autonomously," *Proceeding of the 3rd ACM International Conference on Digital Libraries*, Pittsburgh, PA, pp. 237-246.

Salton, G. (1989), *Automatic text processing: the transformation, analysis, and retrieval of information by computer*, Addison-Wesley, Reading, MA.

Schütze, H. and Silverstein, C. (1997), "A comparison of projections for efficient document clustering," *Proceedings of the 20th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 74-81.

Scott, S. and Matwin, S. (1998), "Text classification using WordNet hypernyms," *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, pp. 38-44.

Sebastiani, F. (2002), "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1-47.

Steinbach, M., Karypis, G. and Kumar, V. (2000), "A comparison of document clustering techniques," *KDD Workshop on Text Mining*, http://citeseer.nj.nec.com/steinbach00comparison.html.

Voorhees, E.M. (1993), "Using WordNet to disambiguate word senses for text retrieval," *Proceedings of the Sixteenth Annual International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 171 – 180.

Weiss, S. and Kasif, S. and Brill, E. (1996), "Text classification in USENET newsgroups: a progress report," *Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access*, Bulgaria, pp. 125-127.

Wermter, S. (1995), *Hybrid Connectionist Natural Language Processing*, Chapman & Hall Neural Computing Series, Chapman & Hall.

Wermter, S. (2000), "Neural network agents for learning semantic text classification," *Information Retrieval*, vol. 3, no. 2, pp. 87-103.

Wermter, S. and Hung, C. (2002), "Selforganising classification on the Reuters news corpus," *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, Taipei, Taiwan, 2002, pp.1086-1092.

Wong, W. and Fu, A. W. (2000), "Incremental document clustering for web page classification," *Proceedings of IEEE 2000 International Conference on Information Society in the 21st Century: Emerging Technologies and New Challenges (IS2000)*, Japan, http://citeseer.nj.nec.com/328087.html.