Wood, S., Fung, C.C. and Gedeon, T. (2003) A hybrid neural network approach for automated classification of online documents using a domain nonspecific thesaurus. In: Fourth International Conference on Intelligent Technologies (InTech'03), 17 - 19 December, Chang Mai, Thailand.

A Hybrid Neural Network Approach for Automated Classification of Online Documents using a Domain Nonspecific Thesaurus

Samea A. Wood
School of Information Technology, Murdoch University
South Street, Murdoch, Western Australia 6150, Australia
swood@murdoch.edu.au

Lance Chun Che Fung
School of Information Technology, Murdoch University
South Street, Murdoch, Western Australia 6150, Australia
L.Fung@murdoch.edu.au

Tamas Gedeon
School of Information Technology, Murdoch University
South Street, Murdoch, Western Australia 6150, Australia
tgedeon@murdoch.edu.au

Abstract: Information overloading has become a serious problem due to the exponential growth of the use of the Internet, emails and other online information resources. One of the solutions to this problem is the deployment of an automated classification system so as to provide an efficient means to manage the ever increasing amount of information and documents. A hybrid neural network approach for the automated classification of text-based articles is reported in this paper. In this study, the research has centered on the classification of newsgroup documents (postings) in accordance to the relevant newsgroups. The classification was initially based on the original documents. The documents are then reclassified with replacement of words from a domanin nonspecific thesaurus. Experiments based on over 40,000 news articles have been carried out and the results are found to be compatible in both cases. The technique can be extended to other online documents such as email articles and web pages.

Keywords Information Retrieval; Document Management, automated classification, thesaurus.

1. Introduction

Exponential growth in the use of online information resources such as email, newgroups and the World-Wide-Web (WWW) has resulted in the availability of an enormous amount of information. This has led to the phemonenon of *information overloading*, or *information overabundance*. Such huge amount of data now hinder effective information management in many organisations. It is obvious that an efficient approach is required to manage the information in a well-organized manner. On the other hand, it is also required to retrieve the information accurately and rapidly as required. *Automated Classification* is a possible solution to the problem. In this paper, the use of a hybrid neural network based approach based on original newsgroup items and use of domain non-specific thesarus are described and the results are analysed.

2. Document Collection

In this study, 40,000 filtered newsgroup postings from ten different newsgroups are collected as test data. The document collection is used to analyse the performance of the proposed hybrid neural network for document classificaiton and retrieval. This source was chosen as the collection of such a large number of documents is easily accessible. In addition, all the items have been pre-classified, with classifications derived automatically from the source of the documents, that is, the news group that the documents belong to.

Ten newsgroups are assumed to be corresponding to ten distinct classes. While the groups are distinctly different, the nature of the document can still be considered similar enough to provide a substantial basis for testing the ability of the system in classifying documents based on different literary genres.

Pre-processing of the document include filtering of advertising posts (spam). In addition, all unwanted punctuation characters and Network News Transport Protocol (NNTP) header information were removed. However, the subject lines are remained intact. The originating 10 newsgroups are listed in Table 1.

Table 1: Ten Newsgroups for testing purpose

Newsgroup	
alt.babylon5.uk	alt.os.windows2000
alt.books	alt.tv.farscape
alt.computer	alt.startrek
alt.movies	rec.humor
alt.os.linux	sci.astro.amateur

The final 40,000 documents were selected randomly from the spam-filtered documents at about

4,000 per newsgroup. Random selection was used in order to obtain a training set representative of the full set of documents. The resultant documents were split into a training and a validation set of 30,000 and 10,000 documents respectively. This results in a combination of 3,000 training and 1,000 testing documents per newsgroup.

The average, minimum and maximum number of lines for each newsgroup document set are shown in Table 2 below.

Table 2: Training & Testing Newsgroup Document Lines

Newsgroup	Average Lines		Min Lines		Max Lines	
	Train	Test	Train	Test	Train	Test
alt.babylon5.uk	39.14	34.73	3	4	788	507
alt.books	44.48	43.78	3	2	3516	3170
alt.computer	32.07	27.54	3	4	746	167
alt.movies	51.12	33.44	3	3	7317	561
alt.os.linux	33.48	36.16	2	1	2949	1121
alt.os.windows 2000	28.23	27.07	3	3	442	350
alt.tv.farscape	36.73	31.61	3	3	178	252
alt.startrek	33.35	31.91	3	4	1034	434
rec.humor	36.21	37.63	3	3	946	716
sci.astro.amateur	31.76	33.29	3	3	1321	444

3. Hybrid Neural Network Approach for automatic document management

An overview of the hybrid neural-network system for text document categorisation is given in Fig. 1 below.

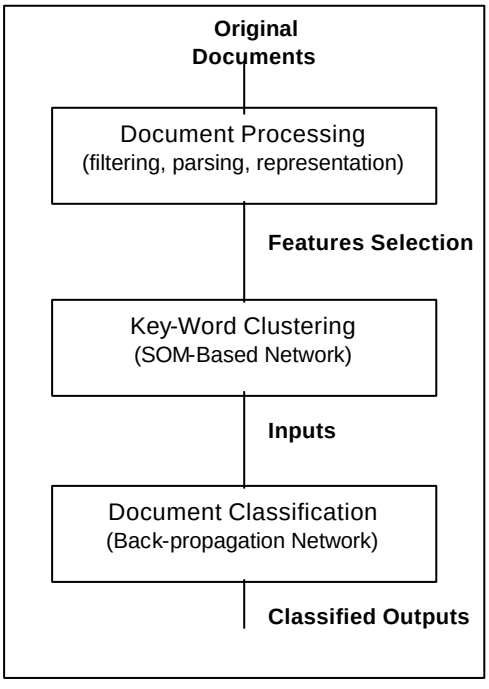

Figure 1: Hybrid Neural-Network Architecture

This hybrid neural-network approach has three main stages. Once document collection and filtering are completed, the first stage involves generating document profiles based on a *document representation* scheme.

The document respresentation scheme involves counting all instances of all words within each document and generating an associative vector of words to word-frequency feature for each document within the training set. These vectors can then be used to determine which words are the most representative of the respective documents as a whole. Based on the word-frequency profiles, it is assumed that this feature will provide sufficient information to represent the diversity and content of the original training documents. For this research, 256 word-frequencies were used to comprise the final document profiles and were selected by examining the difference in total word-frequencies between the ten different document classes. Roughly 25 words per comparison were selected manually, with common words excluded from the result. Words occurring in the newsgroup name were also included, with newsgroup type abbreviations being excluded. The figure of 256 for the number of document-words was selected based on prior research work. While the value of 256 may not be optimal, the largest single factor limiting the number of document-words that could be used is computer processing power.

This representation scheme is enhanced by weighting the significance of word occurance in the document collection. That is, words with high occurance in the document collection are weighted lower than low occurance words. This weighting formula is given below, where w_{ik} is the associated weight, tf_{ik} is the number of occurances of the word

$$w_{ik} = tf_{ik} \times idf_k$$

$$idf_k = \log\left(\frac{N}{n_k}\right)$$

in document i, and id_{fk} is the inverse document frequency of the word t_k in the collection of documents. N is the total number of documents in the collection, of which n_k contain the word t_k.

This scheme is known as *Inverse Document Term Weighting (IDTW)*.

Once the document profiles are generated, they can be used to train the self-organising map at stage 2, which maps the relationships between the

document-words and performs key-word clustering by a Self-Organising Map (SOM).

The SOM consists of 256 input neurons, corresponding to the 256 key-words, and 256 competative layer neurons. From the SOM output, a set of stage 3 training patterns were generated by mapping the highest trained SOM activated neuron index for each document-word, and setting the corresponding stage 3 256 element training pattern feature to "on" if that word occurred in the document. The objective is to provide stage 3 with enough information to allow the classification of documents based on the occurrence and relationships between document-words (mapped by the SOM) for documents in different newsgroups.

Table 3 presents a sample set of SOM key-word clusters.

Table 3: Example SOM Word Clusters

Word	Cluster Index	Word	Cluster Index
ds9	1	jupiter	23
tng	1	saturn	23
alien	31	china	33
aliens	31	chinese	33
delen	88	actress	95
marcus	88	aeryn	95
		claudia	95
captain	66		
kirk	66	silly	170
picard	66	saying	170
ship	66	joke (s)	170

A number of other schemes for mapping the clustered data to stage 3 input patterns were tested, but none has proved to be as effective as 'word-activation mapping'.

Stage 3 involves training a back-propagation (BP) network using the back-error propagation algorithm and the stage 3 training patterns. The purpose of the BP network is to determine the mapping of the document-classification relationships based on document-word cluster knowledge.

The BP networks require supervised training and preclassified data are used. In this case the expected output for the BP network is a 10 element vector with one element for each newsgroup. A single vector element will be turned on representing the newsgroup the original document belongs to.

The BP network architecture consists of 256 input neurons, a single hidden layer optimised through experimentation to 12 neurons, and a 10 neuron output layer.

4. Results of the Hybrid Network for Classification and Recall

4.1. Stage 2 - Self-Organising Map

In order to examine the self-organising map's performance, it is necessary to look at the document-word clustering it generated. Table 3 presents a sample of clusters from one series of SOM word-activation mapping results.

The most apparent attribute of the SOM clustering is the relatively small sizes in some classes. The self-organising map clusters range from 1 word to a maximum of 9 words per cluster. Nonetheless, the majority of clusters that were formed on all SOM training runs were justifiable in the context of the document sets.

4.2. Stage 3 - Back-Propagation Neural Network

Using the standard measures of *recall* and *precision*, a quick examination of the back-propagation network's results (indiciative of the systems performance as a whole) indicates that the system performs best when employing the IDTW document representation scheme. In this instance, recall and precision measures were based on the classification performance of the system, with regard to the documents' sources. A document is correctly classified when the network classification matches the original newsgroup that was the source of the document. These measures were generated for each of the ten classifications of newsgroup documents for both the training and validation sets. The results are shown in Table 4 (a) and (b) respectively.

Given the generally small size of newsgroup documents, and a limit of 256 document words for classifying these documents, many of which had only a small number of document word occurrences, these results are encouraging. Additionally, the interrelated nature of many of the newsgroups further highlights the ability of the system to distinguish between document sets.

Specifically, results show that this hybrid neural network architecture is currently capable of achieving a 63.9% recall a 78.40% precision rate on the validation document set.. While these figures compare favourably with results for nearest neighbourhood clustering of 59.3% recall and 59.9% average precision, and results from a standard back-propagation network trained on the normalised word-frequency vectors of 40.81% recall and 76.45% average, there is clearly room for improvement. The

following section details the inclusion of a domain nonspecific thesaurus to the architecture for the purpose of improving the efficiency of the system.

Table 4: Results of Newsgroup Document Classification

		Recall		Prec
alt.babylon5.uk	2516	83.9%	2568	98.0%
alt.books	2400	80.0%	2547	94.2%
alt.computer	2457	81.9%	2633	93.3%
alt.movies	2443	81.4%	2546	95.9%
alt.os.linux	2570	85.7%	2655	96.8%
alt.os.windows2000	2605	86.8%	2773	93.9%
alt.tv.farscape	2743	91.4%	2776	98.8%
at.startrek	2579	86.0%	2653	97.2%
rec.humor	2257	75.2%	2364	95.5%
sci.astro.amateur	2460	82.0%	2577	95.5%
Average:		**83.4%**		**95.9%**

(a) Training Data Set Results

		Recall		Prec
Alt.babylon5.uk	602	60.2%	756	79.6%
Alt.books	625	62.5%	795	78.6%
Alt.computer	438	43.8%	656	66.8%
Alt.movies	559	55.9%	715	78.2%
Alt.os.linux	743	74.3%	856	86.8%
Alt.os.windows2000	677	67.7%	971	69.7%
Alt.tv.farscape	675	67.5%	785	86.0%
at.startrek	745	74.5%	929	80.2%
rec.humor	640	64.0.%	832	76.9%
Sci.astro.amateur	686	68.6%	845	81.2%
Average:		**63.9%**		**78.4%**

(b) Validation Data Set Results

5. Adding a Domain Nonspecific Thesaurus

It was decided to investigate the possibility of improving the system performance with a domain nonspecific thesaurus. The idea was to replace words which are not key words but having the same meanings of an equivalent keyword or keywords. The integration of a thesaurus is done at stage 1 of the system. At the stage of document representation, the system would allow words that are semantically related to the document-words to be considered during the process of generating document profiles.

For example, given the document-word '*computer*', it may also be desirable to consider the following related words: *adding machine, thinking machine, data processor, electronic brain, laptop, notebook, workstation, supercomputer, mainframe, microcomputer, number cruncher, personal computer and PC.*

By decreasing the specificity of the document-word terms in this manner, it was hoped that a more accurate representational knowledge of the

documents can be included in the existing representation scheme without increasing the number of document-words.

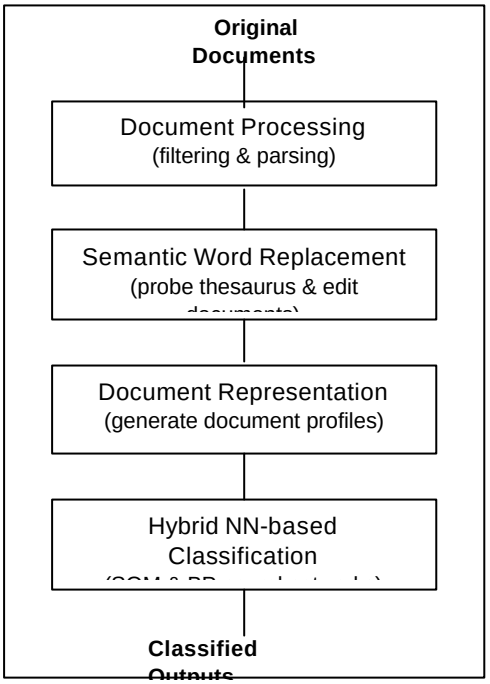

Figure 2: Hybrid Neural-Network Architecture using Semantic Word Replacement

In order to achieve this, all words semantically related to the keywords were replaced by their associated keyword in the document set. This was achieved using an open-source thesaurus application to retrieve the semantically related words, and replacing those words in the original document set. The 'new' document set was then used to train and test the hybrid system in the same manner. The full revised hybrid system is shown in Figure 2.

The results for each of the ten classifications of newsgroup documents for both the training and validation sets using the domain nonspecific thesaurus are shown in Table 5, (a) and (b) respectively.

Table 5: Results using Domain Nonespecific Thesaurus

		Recall		Prec
alt.babylon5.uk	2476	82.5%	2537	97.6%
alt.books	2169	72.3%	2391	90.7%
alt.computer	2455	81.8%	2656	92.4%
alt.movies	2481	82.7%	2698	92.0%
alt.os.linux	2519	84.0%	2621	96.1%
alt.os.windows2000	2580	86.0%	2700	95.6%
alt.tv.farscape	2759	92.0%	2794	98.7%
at.startrek	2541	84.7%	2620	97.0%
rec.humor	2309	77.0%	2564	90.1%
sci.astro.amateur	2288	76.3%	2394	95.6%
Average:		**81.9%**		**94.6%**

(a) Training Data Set Results

		Recall		Prec
alt.babylon5.uk	559	55.9%	728	76.8%
alt.books	524	52.4%	694	75.5%
alt.computer	465	46.5%	713	65.2%
alt.movies	610	61%	811	75.2%
alt.os.linux	758	75.8%	879	86.2%
alt.os.windows2000	650	65%	898	72.4%
alt.tv.farscape	584	58.4%	680	85.9%
at.startrek	660	66%	874	75.5%
rec.humor	655	65.5%	958	68.4%
sci.astro.amateur	595	59.5%	831	71.6%
Average:		**60.6%**		**75.3%**

(b) Validation Data Set Results

6. Discussion and Conclusion

The experiments have shown that the hybrid neural network system has achieved an average recall accuracy of 83.9% for training and 63.9% in the validation testing. The mreasures for precision accounts for 95.9% and 78.4% respectively. On the other hand, the use of the domain nonspecific thesarus has resulted in an 81.9% and 60.6% for recall in the training and validation data sets. The precision measures are 94.6% and 75.3% respectively.

As can be seen from the results presented above, the inclusion of the domain nonspecific thesaurus to the hybrid neural-network system has resulted in a slight decrease in recall and precision of 3.3% and 3.1% respectively. Nevertheless, the results are very similar in both recall and precision measures.

However, it is observed that the results may not be representative of the characteristic classes of documents due to the relatively short length in the newsgroup postings. The inherent lack of keywords, and semantically related words have provided the hybrid system with very little, and in some cases, even null information on which to base its classification. The relatively high precision figures show that despite the systems inability to classify this type of document, it does not usually misclassify them. This demonstrates a useful feature in classifying and retreiving the document using the proposed hybrid neural network. The present study will be extended to other sources such as legal and government documents in order to test the system's suitability and useability.

References

[1] Lam, W., Ruiz, M. and Srinivasan P., (1999) "*Automatic text categorization and its application to text retrieval*" IEEE Trans. on Knowledge and Data Engineering, vol. 11, Issue 6, pp. 865-879.

[2] Gedeon, T.D. and Bustos, R.A. (1996) "*Word-Concept Clusters in Document Collections*", Australian Document Computing Symposium, Melbourne, vol. 1, pp. 21-24.

[3] Kohonen, T., Kaski, S., Lagus, K. and Honkela, T. (1996) "*Very Large Two-Level SOM for The Browsing of Newsgroups*" in the Proceedings of International Conference on Artificial Neural Networks, 1996.

[4] Lin, X. Soergel, D. Marchionini, G. (1991), "*A Self-Organizing Semantic Map for Information Retrieval*", in the Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 262-269.

[5] Troina, G. Walker, N. (1996), "*Document Classification and Searching - A Neural Network Approach"* ESA Bulletin N87, Frascati, Italy.

[6] Wong, SKM., Cai, YJ. and Yao, YY. (1993). "*Computation of Term Association by Neural Network*". SIGIR '93 Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.