Knowledge Based Document Management System for Free-Text Documents Discovery

[1]Paul D Manuel[2], Mostafa Ibrahim Abd-El Barr[3], S. Thamarai Selvi[4]

[2]Department of Information Science, College for Women Kuwait University, Kuwait.
p.manuel@cfw.kuniv.edu

[3]Department of Information Science, College for Women Kuwait University, Kuwait.
Mostafa@cfw.kuniv.edu

[4]Department of Information Technology, Madras Institute of Technology, Anna University, India.
stselvi@annauniv.edu

Abstract

A Knowledge Based Document Management System (KBDMS) is proposed in this paper to organize, cluster, classify and discover free-text documents. Context sensitive information is discovered by means of word map, sentence map and paragraph map in an intelligent manner in this proposed system. A text learning procedure for the semantic retrieval of text documents is implemented using hierarchy of self-organizing maps (SOM) and support vector machines (SVM). The hierarchical SOM generates histograms of paragraph maps based on the semantic similarity and these paragraph maps are trained using SVM for classification. The SVM also generates index for each document given to it. The proposed system is scalable and capable of discovery of documents from a huge amount of free-text documents. It is tested over a maximum of 100000 text documents with $75-80\%$ accuracy in the context-sensitive discovery of free-text documents.

Keywords: knowledge based document management system, context-sensitive discovery of free-text document, self organizing maps, semantic similarity, support vector machines.

AMS Subject Classification: 05C12

1 Introduction

The challenging task of E-Governance is to manage huge volume of documents and to provide intelligent information in a meaningful way. Knowledge Management System is a key task of E-Governance. It is used to organize and cluster documents in the form knowledge libraries so that intelligent reports and precise summary analysis are generated.

[1] This work was supported by Kuwait University, Research Grant No. [WI - 04/05].

Semantic retrieval of information deals with the representation, storage and access to information in a specific context. Information Search is performed by giving keywords to search engines to retrieve target information. But a lot of irrelevant information is also retrieved by keyword searching. Hence there is a need for semantic discovery of information instead of keyword searching.

With the volume of documents growing exponentially, document organizing and clustering become increasingly difficult. As a result, various algorithms for dimensionality reduction in automatic text classification have been developed [12]. The problem has also been addressed by automata theory, grammars and natural language theories. In some applications, rule based systems have also been used. Neural network methods based on competitive learning offer an associative approach to this problem [7]. Scholtes [13] has implemented neuronal methods for free-text database search and concluded that neural network converges towards a proper representation of the query as well as the objects in the database. It is observed that SOM is suitable for organizing free-text documents into meaningful maps for exploration and search. SOM algorithm automatically organizes the documents into a two-dimensional grid so that related documents appear closer to each other. Several SOM based techniques are available mainly for browsing, information retrieval and data mining [8, 9, 10]. The KBDMS is implemented using SOM hierarchy, coding schemes and Multi Layer Perceptions (MLP) for dimensionality Reduction, creation of indexable document feature maps respectively.

This paper aims at introducing techniques for discovering relevant text-based data in database of electronic documents using semantic similarity. An electronic database is typically a set of documents in which each data is represented by keywords, sentences or paragraphs available in the given text. The documents most related to the searching query are based on the frequency of occurrences of words, sentences and paragraphs. However, it is essential to extract these details from the free-text document, which are clustered and organized in neural networks for later discovery. As a result, the system provides a way to represent the documents based on their content.

The paper is arranged as follows. The Related Work is provided in Section 2; the Proposed Architecture is discussed in Section 3; the Design and Implementation are given in Section 4; the results are given in Section 5 and Section 6 concludes the paper.

2 Related Work

Kohonen [3] constructed vectors as weighted word histograms using SOM based on shortcut computational methods resulting in content addressable search. The accuracy of search has been increased to 64% from 58% in this approach. Marie-Jeanne [4] proposed a kernel based methodology defining a hierarchical kernel for SOM which compares paragraphs of free-text documents. Topographic clustering has been formed using kernel based learning algorithm in SOM. The inter-cluster distance has been used as the metric for semantic similarity. It has been concluded that the method is effective in retrieving the relevant information.

Andreas Rauber [5] explored the high potential of the SOM for document clustering without intellectual input. The automatically produced labels yield the important keywords describing the contents of the documents. Kin Keung Lai [6] described the use of SOM as a meta-modeling technique to design a parallel text data extrapolation system. He has claimed that the computational efficiency and scalability of the meta-modeling system have been improved when applied to a massive text data collection. Xiaohua Zhhou et al., [15] proposed the agglomerative clustering for grouping documents based on semantic similarity using Vector cosine and Kullback-Leibler divergence metrics. He has concluded that Kullback-Leibler divergence distance metric outperforms Vector cosine metric in finding semantic similarity between the documents. Xiaodan Zhang et al., [16] evaluated the effect of semantic smoothing with a model-based agglomerative clustering and model-based partitioned clustering on three different datasets. They concluded that the proposed semantic smoothing is very promising for model based text clustering. Ari Visa et al., [1] proposed Knowledge Discovery from Text Documents Based on Paragraph Maps. We have made use of this concept in introducing an efficient document management system by indexing the documents using back propagation in addition to hierarchical SOM.

3 Proposed Architecture

A framework for KBDMS proposed. It consists of three-layer architecture: the Presentation layer, the Knowledge layer, and the Data layer (see Figure 1). The functionality of each layer is discussed below.

3.1 Presentation Layer

The presentation layer provides the user interface to KBDMS. It provides facilities to interact with the training and discovery modules.

3.2 Knowledge Layer

The knowledge layer is the vital part of KBDMS. Since it is implemented in a generic way, knowledge can be introduced through Machine Learning (which is considered in our work), artificial intelligent techniques and statistical training methods or by any soft computing methodology. Further, it can act intelligently through the inclusion of modules such as pattern recognition, logical reasoner, or data mining models. The knowledge layer enables intelligent interpretation of the existing document by extracting meaningful features for decision making as shown in the following algorithm.

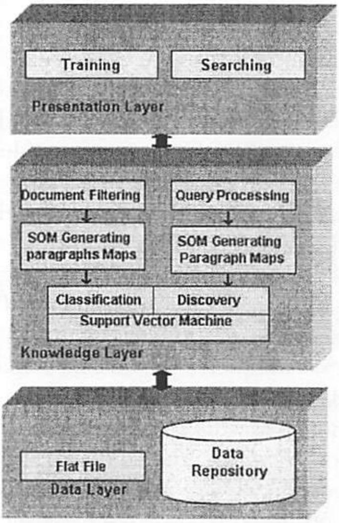


Figure 1: Proposed Architecture for Knowledge Based Document Management System

Algorithm:

Training Session

Step 1: Stop-words and special symbols are removed.

Step 2: Generate word map followed by sentence map and then paragraph map with the help of filtered text from Step 1.

Step 3: The generated paragraph maps are given as input to the SVM which creates the unique index.

Step 4: Store document along with the generated index in the database.

Discovery Session

Step 1: Preprocess the query string; stop-words and special symbols are removed.

Step 2: Generate word map followed by sentence map and then paragraph map with the help of filtered text from Step 1

Step 3: Provide the generated paragraph map to a trained Neural Network Model to discover the relevant document from the database.

3.3 Data Layer

This layer is used for storing the data permanently in repositories or in the form of flat file. The data may also be stored in any other required structure.

4 Design and Implementation

In this section, we explain the design and the implementation details of the constituent units of each layer viz. presentation layer, knowledge layer and data layer.

The presentation layer is implemented using Java swing. The GUI screen (see Figure 2) allows user to either select the training or discovery. If the user selects to train the KBDMS, he/she can enter the filename of the training document in the text field. On the other hand, if the user wants to perform discovery of documents, he/she enters the search query through the text field available under discovery section. On clicking the appropriate buttons, training or discovery in the presentation layer, the knowledge layer for further processing is invoked.

The knowledge layer is implemented using MATLAB Neural Network Toolbox. The SOM used in this layer is a variant of the Kohonen [2]] model which uses the Kullback-Leibler metric instead of the Euclidean distance metric. Given two probabilistic document models $p(w \mid d_1)$ and $p(w \mid d_2)$, the Kullback-Leibler divergence distance of $p(w \mid d_1)$ to $p(w \mid d_2)$ is defined as:

$$\Delta(d_1, d_2) = \sum_{w \in V} p(w \mid d_1) \log \frac{p(w \mid d_1)}{p(w \mid d_2)} \tag{1}$$

where V is the vocabulary of the corpus.

The document filtering unit removes the stop-words and special symbols. The filtered text is provided to the hierarchical SOM to generate paragraph maps. Initially, words in the filtered text are converted into numbers using Ari Visa et al. [1] formula as given below.

$$y = \sum_{i=0}^{L-1} 2^{4i} * C_{L-i} \tag{2}$$

where L is the number of characters in the word C_i and i is a character within a word w.

In the next step, the corresponding numbers of words are further mapped into word vectors using the formula

$$Y = f(y) \bmod P \tag{3}$$

Here Y is a tabulated function, the table has a size of $N * M$ where N is the length of the table and M is the length of the word vector, P is a suitable prime and $N < P < NW$.

Finally the constructed word vector is given to the SOM which forms the word map. Similarly, the sentence vector is created using the filtered text and word map. The sentence vector generated is fed to the SOM to form sentence map.

The paragraph vector is generated by the SOM using the word map, sentence map and filtered text. The paragraph map is also generated by the SOM. The classification module uses SVM, in which the paragraph map is trained and a corresponding index is generated. The indices of the trained documents are stored in repository for subsequent discovery.

In the discovery process, the creation of index for the search string follows the same steps explained above. Based on the generated index of the search string, the suitable matching database entry is returned as the result for the discovery operation.

The data layer enables the persistent storage of the indices generated by the Knowledge Layer. The salient feature of data layer is the ability to store indices along with the document in repository. This layer may also be used to store document as a flat file with the same name of the generated index. Thus the data layer is extensible to accommodate the required structure of the document.

The Support Vector Machine is used for classification. The generalization ability of SVM is better than other neural classifiers. The SVM is implemented using LS-SVM library files. The Java front-end design is connected to the MATLAB tool. The steps involved in connecting Java with MATLAB are given below:

1. Importing Packages
2. Registering MATLAB Engine
3. Opening a Connection to MATLAB
4. Executing Result Set.

A complete document discovery system has been implemented in a generic way using semantic similarity and employing SOM and SVM.

5 Results & Discussion

The KBDMS is trained with data sets in the range of 50000, 80000 and 100000 text documents. The training time ranges from 8 hours to 15 hours in a Pentium IV machine. The accuracy of the context sensitive discovery of documents has been observed in the range of 75% to 80%.

Initially a training set containing 900 words is given to the KBDMS. The training set is filtered and the subsequent encoding of the filtered text gives the word vectors, which are given to the SOM that generates the word map. The word map and the filtered text are given to the sentence encoding that produces the sentence vectors, which in turn is fed to the SOM to get the sentence map. The word map and the sentence map obtained are plotted as histogram as in Figure 3. The word map, sentence map and filtered text are fed to the paragraph encoding to get the sentence vectors. The sentence vector is input to the SOM to get the sentence map. The sample sentence map is represented as histogram in Figure 4. Figure 5 shows the learning trial of the SVM for the given training set. The effective retrieval of the related Free-Text Document for the given query consisting of 10 words is shown in Figure 6. Figure 7 shows a plot to show the multi-class classification ability of LS SVM.

6 Conclusion

A generalized framework for knowledge based document system is proposed in this paper. The model is implemented using hierarchical SOM and SVM for efficient discovery of free-text document. The contributions in this paper are:

1. A generalized knowledge based document management system is proposed.
2. New algorithms are designed using Kullback-Leibler divergence metric for clustering the document in the SOM.
3. Knowledge discovery is achieved through SVM by indexing the documents.

By using SOM and SVM for discovery and considering KLD for semantic similarity the efficiency is improved and the accuracy achieved was nearly 10% higher than the other existing methods. The hierarchical SOM is used for generating word, sentence, and paragraph maps. Free text documents are trained and indexed for fast recovery, while the existing methods do not apply any indexing mechanism. The proposed approach has been implemented and the results are found to be satisfactory and encouraging.

Acknowledgement: This work was supported by Kuwait University, Research Grant No. [WI - 04/05].

References

[1] Ari Visa, Jarmo Toivonen, Piia Ruokonen Hannu Vanharanta and Barbro Back. "Knowledge Discovery from Text Documents Based on Paragraph Maps". Proceedings of the 33rd Hawaii International Conference on System Sciences – 2000.

[2] Kohonen, Self-Organizing Maps, Springer Verlag, New York, Inc. 2001.

[3] Teuvo Kohonen, Samuel Kaski, Krista Lagus, Jarkko Salojarvi, Jukka Honkela, Vesa Paatero, and Antti Saarela, "Self Organizing of Massive Document Collection". IEEE Transaction on Neural Networks, Vol 11, No 3, May 2000.

[4] Marie-Jeanne Lesot, Delphine Dard and Florenced'Alché-Buc, "A Methodology for Topographic Clustering of Structured Text Documents". International conference on Learning Methods for Text Understanding and Mining, 26 - 29 January 2004, Grenoble, France.

[5] Andreas Rauber, Erich Schweighofer, Dieter Merkl, "Text Classification and Labeling of Document Clusters with Self-Organising Maps". OEGAI Journal, 19:17—23, 2000.

[6] Kin Keung Lai, Lean Yu Ligang Zhou, and Shouyang Wang, "Self-Organizing-Map-Based Meta-modeling for Massive Text Data Exploration". LNCS 3971, pp. 1261 - 1266, 2006.© Springer-Verlag Berlin Heidelberg 2006.

[7] Kohonen T, "Self-Organized formation of topologically correct feature maps". Biological Cybernetics, Vol 43, PP 59–69, 1982.

[8] Ultsch A, "Knowledge Acquisition with Self-Organizing Neural Networks". Artificial Neural Networks, 2, Volume I, Pages 735-738, Amsterdam, Netherlands, 1992, North-Holland.

[9] Kohonen T., S.Kaski, K.Lagus, and T.Honkela, "Very Large Two-Level SOM for Browsing of Newsgroups". In Proc. of ICANN'96. International Conference on Artificial Neural Networks, Pages 269-274. Springer, 1996.

[10] Landauer T.K.and S.T.Dumais, "A solution to plato's problem: The Latent Semantic analysis theory of acquisition, induction and representation of knowledge". Psychological Review, 104:211-240, 1997.

[11] Lin X, D.Soergel, and G.Marchionini, "A Self-Organizing Semantic Map for Information Retrieval". 14th International ACM/SIGIR Conf. On R&D in Information Retrieval, Pages 262-269, 1991.

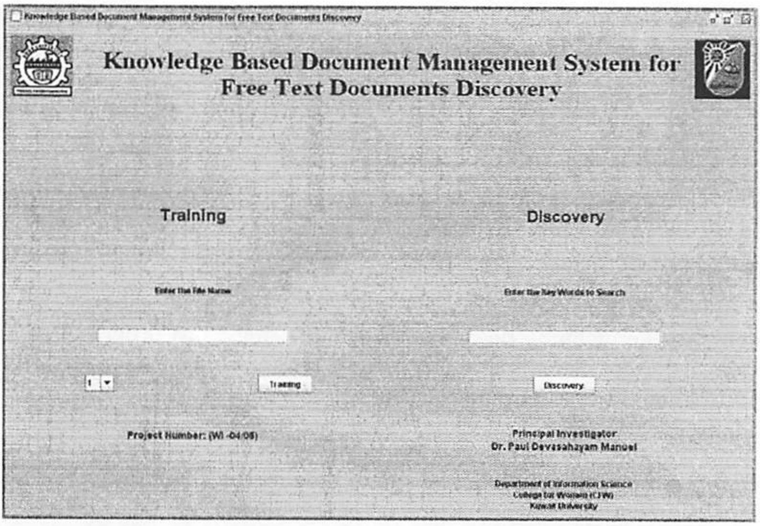
[12] Christos Faloutsos and King-Ip Lin.FastMap(1995), "A Fast Algorithm for indexing, data-mining and visualization of traditional and multimedia datasets". Poceedings of the ACM SIGMOD International Conference on Management of Data. Pages 163-174, San Jose, California, 22-25,1995.

[13] Scholtes J.C, "Unsupervised Learning and the information retrieval problem". Int. Joint Conference on Neural Networks - IJCNN'91, Volume I, Pages 95-100, Piscataway, NJ, 1991.

[14] Landauer T.K., D.Laham, R.Rehder, and M. E. Schreiner, "How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans". In Proceedings of the 19th Annual Conference of the Cognitive Science Society, Pages 412-417, Mahwah, NJ, 1997.

[15] Xiaohua Zhou, Xiaodan Zhang and Xiaohu Hu, "Semantic Smoothing of Document Models for Agglomerative Clustering". Twentieth International Joint Conference on Artificial Intelligence (IJCAI 2007), Jan. 6-12, 2007, India,2922-2927.

[16] Xiaodan Zhang, Xiaohua Zhou and Xiaohu Hu, "Semantic Smoothing for Model-Based Document Clustering". IEEE International Conference on Data Mining (IEEE ICDM06), Dec. 18-22, 2006, Hong Kong, pages 1193-1198.



Knowledge Based Document Management System for Free Text Documents Discovery

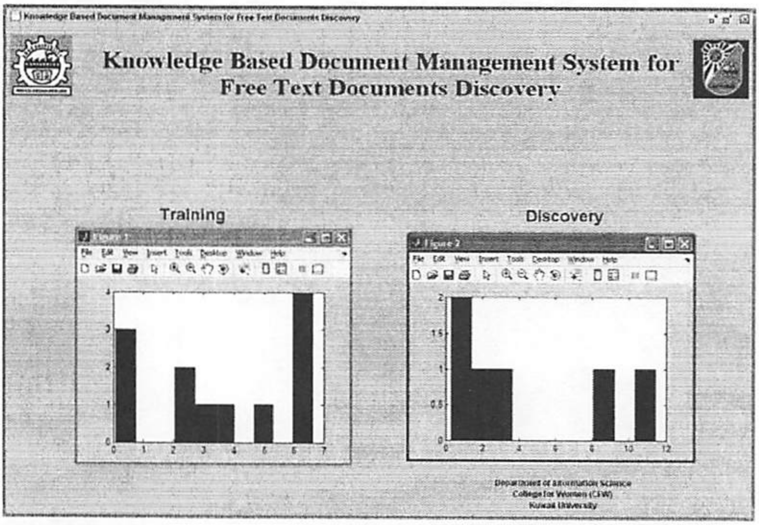
Knowledge Based Document Management System for Free Text Documents Discovery

Project Number: (WI -04/05)

Principal Investigator
Dr. Paul Devasahayam Manuel

Department of Information Science
College for Women (CFW)
Kuwait University

Figure 2: Graphical User Interface



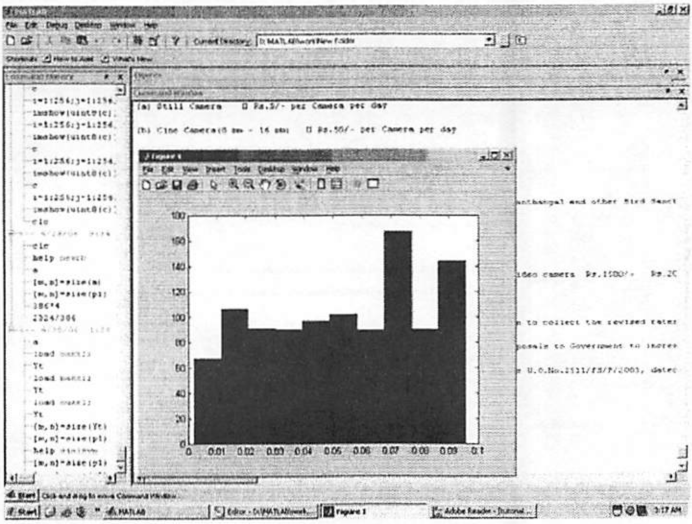
Knowledge Based Document Management System for Free Text Documents Discovery

Knowledge Based Document Management System for Free Text Documents Discovery

Department of Information Science
College for Women (CFW)
Kuwait University

Training
Discovery
Figure 1
File Edit View Insert Tools Desktop Window Help
4
3
2
1
0
0 1 2 3 4 5 6 7
Figure 2
File Edit View Insert Tools Desktop Window Help
2
1.5
1
0.5
0
0 2 4 6 8 10 12

Figure 3: Plot to show the word map and sentence map during the training procedure of Free-Text Documents

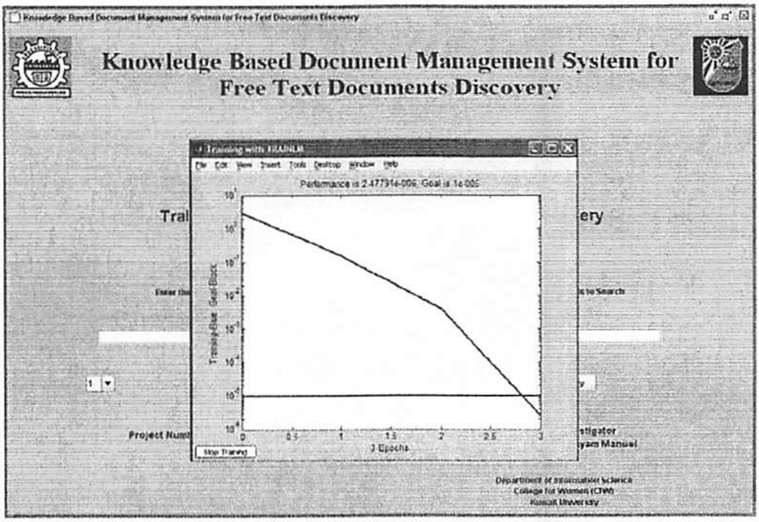


MATLAB
File Edit Debug Desktop Window Help
Current Directory: D:\MATLAB\work\New Folder
Shortcuts How to Add What's New
Command History
Command Window
Figure 1
File Edit View Insert Tools Desktop Window Help
100
160
140
120
100
80
60
40
20
0
0 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.1
Start Click and drag to move Command Window
Start MATLAB Editor - D:\MATLAB\work... Figure 1 Adobe Reader - [tutorial...]

```
c
i=1:256;j=1:256;
imshow(uint8(c));
i=1:256;j=1:256;
imshow(uint8(c));
c
i=1:256;j=1:256;
imshow(uint8(c));
c
i=1:256;j=1:256;
imshow(uint8(c));
clc
clc
help [illegible]
a
[m,n]=size(a)
[m,n]=size(p1)
386*4
2324/386
a
load [illegible]
Yt
load [illegible]
Yt
load [illegible]
Yt
[m,n]=size(Yt)
[m,n]=size(p1)
help [illegible]
[m,n]=size(p1)
```

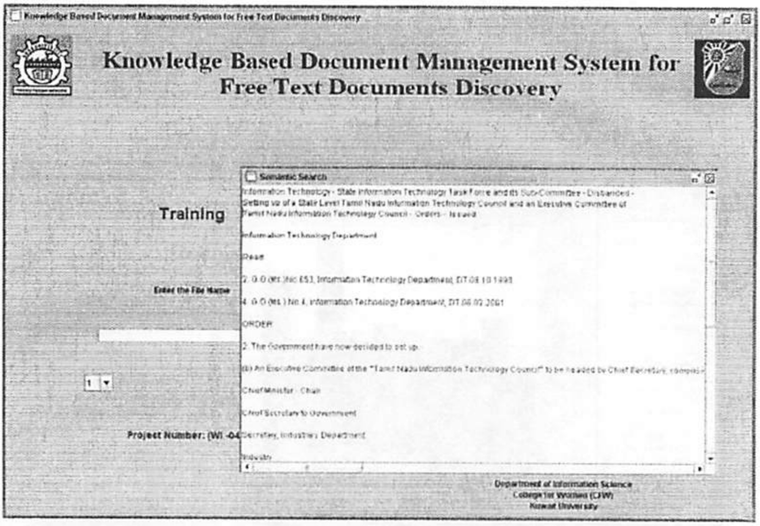
```
(a) Still Camera    □ Rs.5/- per Camera per day

(b) Cine Camera(8 mm - 16 mm)    □ Rs.50/- per Camera per day
```

Figure 4: Plot to show the Sentence histogram by using Knowledge maps.



Knowledge Based Document Management System for Free Text Documents Discovery
Knowledge Based Document Management System for
Free Text Documents Discovery
Training with TRAINLM
File Edit View Insert Tools Desktop Window Help
Performance is 2.47791e-006, Goal is 1e-005
Training-Blue Goal-Black
3 Epochs
0 0.5 1 1.5 2 2.5 3
Stop Training
Trai
ery
s to Search
Project Num
stigator
yam Manuel
Department of Information Science
College for Women (CFW)
Kuwait University

Figure 5: Plot to show the learning trial of the SVM.



Knowledge Based Document Management System for Free Text Documents Discovery

Knowledge Based Document Management System for Free Text Documents Discovery

Training

Enter the File Name

1

Project Number: (WI -04

Semantic Search

Information Technology - State Information Technology Task Force and its Sub-Committee - Disbanded - Setting up of a State Level Tamil Nadu Information Technology Council and an Executive Committee of Tamil Nadu Information Technology Council - Orders - Issued

Information Technology Department

Read:

2. G.O.(Ms.)No.653, Information Technology Department, DT.08.10.1998

4. G.O.(Ms.) No.4, Information Technology Department, DT.06.02.2001

ORDER:

2. The Government have now decided to set up:

(b) An Executive Committee of the "Tamil Nadu Information Technology Council" to be headed by Chief Secretary, comprisi

Chief Minister - Chair

Chief Secretary to Government

Secretary, Industries Department

Industry

Department of Information Science
College for Women (CFW)
Kuwait University

Figure 6: Plot to show the effective retrieval of the related Free-Text Document for the given query.

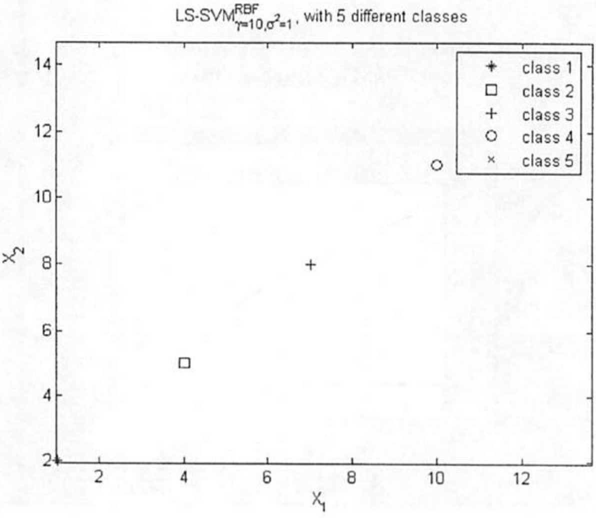


Figure 7: Plot to show the multi-class classification ability of LS SVM.