Comparison of IR Models for Text Classification

Priyanka Desai
PhD research Scholar
JJTU
Rajasthan, India

G.R.Kulkarni, PhD
Principal
R.W.M.T'S DNYANSHREE INSTITUTE OF ENGINEERING AND TECHNOLOGY
Satara, Maharashtra

ABSTRACT
As there is availability of large amount of data on the web, but due to constraints web is only used for browsing and searching. Traditional IE uses NLP techniques such as lexicons, grammars, whereas web applies machine learning and pattern mining techniques to exploit the syntactical patterns or layout structures of the template-based documents.Information Retrieval is the art of presentation, storage, organization of and access to information items.IR now-–days mainly deals with retrieving information based on user queries. The paper deals with basic understanding of IR and IR models and shows Support Vector Machines is a good technique fir classification of huge data sets.

General Terms
Information Retrieval (IR)

Keywords
Boolean retrieval,Vector Space model, BM25, Language models, Inference networks, Learning to Rank and SVM

INTRODUCTION
The definition of information retrieval according to (Manning et al., 2009): (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers) [1].The first process in IR is to index language which is used to describe documents and queries for keyword search where searching is used for controlling vocabularies. The other processes are to remove stop words,stem the words. Edit distance is used to transform one string to another.Soundex is used for phonetic matching. The Index term weighting is related to number of index terms assigned to a given document. In Index term weighting Zipf's law is used for different languages when indexing and Luhn's analysis is used for resolving power of words.The tf-idf shows how important a word is to a document in a whole collection of documents. The tf-idf weight: where weight of a term is the product of its tf weight and its idf weight.$W_{t,d}$= (1+log $tf_{t,d}$).logN/df_t, tf is the terms present in all documents, idf gives importance to few important documents is the total number of documents available,$tf_{t,d}$ of term t in document d is defined as the number of times that t occurs in d, dftis the number of terms. Finally as shown in Figure 1, the Inverted file is a word oriented mechanism which indexes collection to speed up searching. The inverted file for retrieving number of documents uses Boolean retrieval, for document number and term weight (tf-idf) uses ranked retrieval and for word offsets of each occurrence of the term uses proximity operators. The feedback mechanism will help find better results.

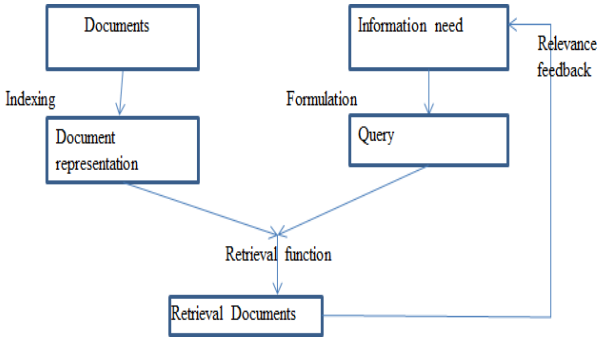

Figure 1:Conceptual IR model [3]

1. PROBLEM WITH IR
The web is growing fast as a result more and more data is available online .As compared to traditional data retrieval which is structured, IR faces a lot of issues as the data is unstructured as shown in Figure 2 and availability of totally different datasets (bulk documents, dynamism of the Internet, duplication, heterogeneity, high linkage, ill-formed quires, etc…). Structured data follows a predefined rule which can be stored in rows and columns such as Relational Data Base Management System, Product (make, manufacturer, year of manufacture etc.). Unstructured data cannot be stored in rows and columns such as the HTML page which contains images, links etc. This unstructured data has to be classified and categorized so as to retrieve/extract data from the web based on the input/query given by the user. There are three primary functions in IR: Indexing- is a process of creating useful index for documents, Search request- has to create query that should retrieve information that is relevant for the user, Request document matching- deals with comparing the created index with formulated request from the user. The challenge is to meet the need of the user to retrieve data from unstructured data. The representation and organization of information

should be in such a way that the user can access information to meet his information need. The paper goes on to explain background of IR, IR models.

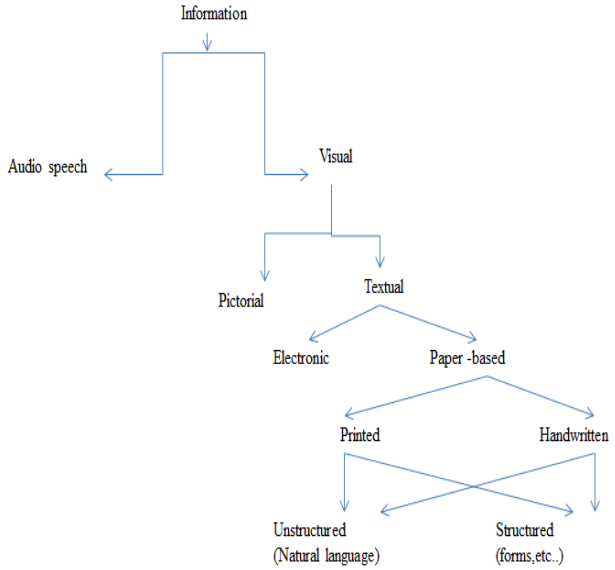

Figure 2 :Various forms in which information is available[2]

2. EXISTING IR MODELS:

IR is used for storing and managing information of documents .IR Models act as guide when retrieving information. For better predications models should have intuition and metaphor.

2.1 Traditional models

2.1.1 Boolean retrieval

In Boolean retrieval there are two possible outcomes for query processing i.e., TRUE and FALSE given by R(q,d)=1 if d->q, 0 otherwise .Queries are usually specified using Boolean operators such as AND, OR, NOT.Following example gives the sequence of queries driven by number of retrieved documents.

Example [4] "Lincoln" searches of news articles

1. Lincoln
2. President AND Lincoln
3. President AND Lincoln AND NOT (automobile OR car)
4. President AND Lincoln AND biography AND life AND birthplace AND Gettysburg AND NOT (automobile OR car)
5. President AND Lincoln AND (biography OR life OR birthplace OR Gettysburg) AND NOT (automobile OR car)

2.1.2 Vector Space model

In Vector Spacemodel each document or query is represented as a vector of term weights. A collection of documents is represented by a matrix of term weights.

Example [5]: Query: "tropical fish

D_1 Tropical freshwater aquarium fish.

D_2 Tropical fish, Aquariumcare, tank setup.

D_3 Keeping tropical fish and goldfish in aquariums and fish bowls.

D_4 the tropical tank homepage-tropical fish and aquariums.

The Figure 3 gives the table for vector representation of stemmed documents without stop words.

Terms	Documents			
	D_1	D_2	D_3	D_4
Aquarium	1	1	1	1
Bowl	0	0	1	0
Care	0	1	0	0
Fish	1	1	2	1
Freshwater	1	0	0	0
Goldfish	0	0	1	0
Homepage	0	0	0	1
Keep	0	0	1	0
Setup	0	1	0	0
Tank	0	1	0	1
Tropical	1	1	1	2

Figure 3: Vector Representation of Stemmed words

VSM ranking function is called retrieval status value (RSV) which gives the cosine similarity between query and document the equation is given as follows

$$R(q,d) = RSV(q,d) = \frac{\sum_{i=1}^{|v|} d_i \cdot q_i}{\left(\sum_{i=1}^{|v|}\sqrt{d_i^2}\right)\left(\sum_{i=1}^{|v|}\sqrt{q_i^2}\right)}$$

$$= \cos\alpha$$

$$= SIM(\vec{q},\vec{d}) = \cos\alpha$$

Example: Consider two documents D_1 and D_2, and a query Q, D_1 = (0.5, 0.8, 0.3), $D2$ = (0.9, 0.4, 0.2), Q = (1.5, 1.0, 0)

$$\cos(q, d_1) = \frac{(0.5 * 1.5) + (0.8 * 1.0)}{\sqrt{(0.5^2 + 0.8^2 + 0.3^2)(1.5^2 + 1.0^2)}}$$

=0.87

$$\cos(q, d_2) = \frac{(0.9 * 1.5) + (0.4 * 1.0)}{\sqrt{(0.9^2 + 0.4^2 + 0.2^2)(1.5^2 + 1.0^2)}}$$

=0.97

q_i is tf-idf weight of term i in the query

di is the tf-idf weight of term i in the document

2.2 Probabilistic Models

"Given a user queryq and a documentd, estimate the probability that the user will findd relevant documents."TheBinary independence retrieval model (BIRM) is one of the important pieces of IR theory. A ranking based on the probability of relevance is optimal with respectto a cost function where the costs for reading relevant documents arelow and the costs for reading non-relevant documents are high(probability ranking principle).Rank the probability of a document which is relevant to the query P(r|q,d) where documents are represented as binary term vector where estimation of P(r|q,d)cannot be done directly ,hence we use Baye's rule to obtain P(d|q,r) which leads to the function g(q,d).

2.2.1 BM25model

BM25 is used by search engines to rank matching documents to their relevance as a ranking function to a given query.Ranks a set of documents based on the query terms in each document,regardless of the inter-relationship modern full-text search collection ,model should pay attention to term frequency and document length[4][5].

The simplest score for document d is idf weighting of the query terms present in the document.

$$\mathrm{RSV_d} = \sum_{t \in q} \log \frac{N}{df_t}$$

Improve the idf term by factoring in term frequency and document length.

$$\mathrm{RSV_d} = \sum_{t \in q} \log \frac{N}{df_t} . \frac{(k_1 + 1).\, tf_{td}}{k_1 \left((1 - b) + b. \left(\frac{L_d}{L_{ave}} \right) \right) + tf_{td}}$$

$tf_{td:}$ term frequency in document d

L_d,L_{ave} length of document d (average document length in the whole collection).

k_1: tuning parameter controlling the document term

b: tuning parameter controlling the scaling by document length

Example:Document=d1=abcbd,d2=befb,d3=bgcd,d4=bde,d5=abeg,d6=bghh,,Word a b c d e f g h ,df_t =2 6 2 3 3 1 3 1,N=6,Query=q=ach, Assume k=1 and b=0.5

$$\mathrm{RSV_{d6}} = \sum_{t \in q} \log \frac{6 + 1}{1 + 0.5} . \frac{(1 + 1).\, 2}{2 \left((1 - 0.5) + 0.5. \left(\frac{4}{4} \right) \right) + 1}$$

Similarly calculate for $\mathrm{RSV_{d1}}$,$\mathrm{RSV_{d2}}$,$\mathrm{RSV_{d3}}$,$\mathrm{RSV_{d4}}$,$\mathrm{RSV_{d5}}$ and get their average.

For long queries, use similar weighting for query terms.$\mathrm{RSV_d} = \sum_{t \in q} \log \frac{N}{df_t} \frac{(k_1+1).tf_{td}}{k_1 \left((1-b) + b \frac{L_d}{L_{ave}} \right) + tf_{td}} . \frac{(k_3+1).tf_{tq}}{(k_3+1).tf_{tq}}$

tf_{tq}: term frequency in query q

k_3: tuning parameter controlling term frequency scaling of the query

k_1 and k_3 values to be set between 1.2 and 2 and b=0.75. BM25 is an approximation to 2 poission process.

For state of art ranking use BM25 or Language models.

2.2.2 Language model

According to [Zhai 04]:LM, which are based on *statistical theory* and *natural language processing* (NLP), have been successfully applied to the problem of ad-hoc retrieval.LM approaches *estimate* a LM for each document and then *rank* documents by the likelihood of the query according to the estimated LM[6].

2.2.2.1 Types of LM's

Build probabilities over sequences of terms, P ($t_1t_2t_3t_4$) =P (t_1) P ($t_2|t_1$) P ($t_3|t_1t_2$) P ($t_4|t_1t_2t_3$)

Unigram language model (the simplest form of LM)

The Probability distribution of words in a language generates the text which consists of pulling words out of a "bucket" according to the probability distribution and replacing them.

$$M \rightarrow P_{uni}(t_1t_2t_3t_4) = P(t_1)P(t_2)P(t_3)P(t_4)$$

where r <-red,b<-blue,y<-yellow

Example->r b r , b y b, r r y,Query is r y r b

Hence

P(r y r b| r b r, b y b, r r y) =4/9∗2/9∗4/9∗3/9 = 96/6561

P(r y r b|b r y,b y b,r r y)=3/9∗3/9∗3/9∗3/9 = 81/6561

P(r y r b|r b b,b b y,r y y)=2/9∗3/9∗2/9∗4/9 = 48/6561

P(r y r b|r y y,r y y,b b y)=2/9∗5/9∗2/9∗2/9 = 40/6561

N-gram language model

Some applications (such as speech recognition) use *bigram* $P_{bi}(t_1t_2t_3t_4)$= $P(t_1)P(t_2|t_1)P(t_3|t_2)P(t_4|t_3)$ and *trigram* language models where *probabilities* depend on previous words. Generate a similar example of probability solutions as in Unigram LM for bigram ,trigram and N-gram models.

2.2.3 Relevance Feedback(rocchio algorithm)

A query can be written in many ways. Rocchio algorithm tries to minimize these variations by using relevance feedback[1][7].

,Cr is the set of relevant documents and

Cnr is the set of non relevant documents, then we wish to find

$$\vec{q}_{opt} = \arg\max \left[\mathrm{sim}(\vec{q}, \vec{C}_r) - \mathrm{sim}(\vec{q}, \vec{C}_{nr}), \vec{q} \right.$$

sim is defined under cosine similarity, the optimal query vector, $\vec{q}_{opt}$ for separating the relevant and non relevant documents:

$$\overrightarrow{q_{opt}} = \frac{1}{|C_r|} \sum_{\vec{d}_j \in C_r} \vec{d}_j - \frac{1}{|C_{nr}|} \sum_{\vec{d}_j \in C_{nr}} \vec{d}_j$$

The optimal query is the vector difference between the centroids of the relevant and non relevant document.. This observation is not terribly useful, precisely because the full set of relevant documents is not known.

Rocchio 1971 Algorithm (SMART) used in practice by making additional assumptions relevance feedback in VSM for $\vec{q}_{opt}$ we rewrite as$\vec{q}_m$

$$\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \epsilon D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \epsilon D_{nr}} \vec{d}_j$$

$\vec{q}_0$ is the original query

$\frac{1}{|D_r|}$are documents which are wanted

$\frac{1}{|D_{nr}|}$documents which are unwanted

α ,β ,γ are weights

If there are a lot of judged documents, there needs to be higher β and some of the weights can be negative which needs to be ignored(set to 0),positive feedback is always more valuable so, set γ < β; e.g. γ = 0.25, β = 0.75,if (γ=0) the system has positive feedback. A good relevance feedback improves recall and precision.

2.3 Combining evidence

Combine different potential documents for effective information retrieval. The potential documents can be simple word based,structure based, PageRank based, metadata based evidence, even scores from different models.

2.3.1 Inference networksmodel

Inference networks is one of the approach to combining evidence which uses Bayesian inference network formula where evidence about a document's relevance to a query is combined from different sources to produce a final relevance judgment. Inference network proposed by Turtle, where models are inferred as Bayesiannetwork. Combines evidence from multiple document representations and document identifiers. Complex information needs can be easily expressed by rich structured query language. Basic model used tf.idf estimation,the new models use the LM estimation. Exact inference match is used for automated systems when there is no user interaction[8][9][10].

Example: Exact inference ,given a Bayesian network and a random variable X,P(X=x)>0,this will work for all network configurations. The diagram is represented in Figure 4.

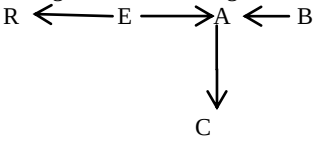

Figure 4:Exact inference of random variables

$$P\left(\frac{B}{C=1}\right)=\frac{P(B,C=1)}{P(C=1)}$$

2.3.2 Learning to Rank

Learning to rank for Information Retrieval (IR) is a task to automatically construct a ranking model using training data, such that the model can sort new objects according to their degrees of relevance, preference, or importance. There can be a variety of learning to rank applications such as Information Retrieval (IR), Natural Language Processing (NLP), and Data Mining (DM).Document retrieval is shown in the Figure 5. Given a query the document is ranked and retrieved, ranking model f(q,d) is used to sort the documents, where q denotes a query and d denotes a document.

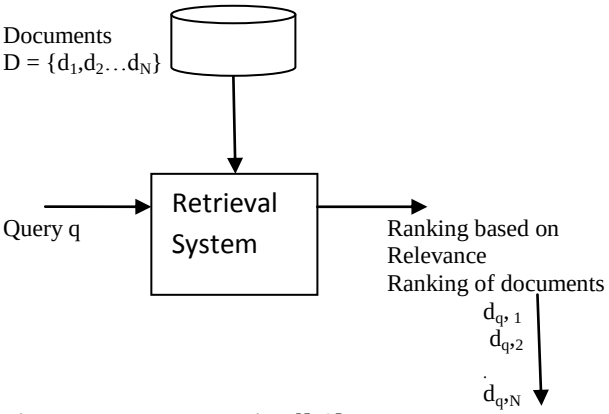

Figure 5:Document Retrieval[13]

3. SOLUTION TO THE IR PROBLEM

As there is a lot of data available on the web automatic text classification is the need of IR. Feature selection and transformation is performed on the documents to be easily understood by Machine Learning techniques. Naïve bayes is generally used because it is simple to implement but does model well for text. SVMs were introduced in COLT-92 by Boser, Guyon &Vapnik. In their basic form SVMs learn linear threshold function.Support Vector Machines(SVM) gives good precision but bad recall for text classification[23], for improving recall adjust the threshold value.

For better classification the proposed method is Support Vector Machines which uses hyper planes which is a good technique for training data. Support vector machines are used for classifying huge amounts of high dimensional data into different classes.Get the score of documents to optimize by sorting the document ,W^TX,where is the hyperplane ,X is the set of documents.A good solution would be to classify the text in proper manner by considering a threshold value for each classification before extracting to have structured data.

By giving some training data D, a set of n points of the form

$D=\{(x_i,y_i)|\ x_i \epsilon R^p, y_i \epsilon (-1,1)\} i=1 \text{ to } n$

where yi is either 1 or −1, indicating the class to which the point X_i belongs. Each X_i is a p-dimensional real vector. To find the maximum-margin hyperplane that divides the points having $y_i = 1$ from those having $y_i = -1$. Any hyperplane can be written as the set of points X satisfying W.X-b=0where (·) denotes the dot product and w the normal vector to the hyperplane. The parameter b/|| W || determines the offset of the hyperplane from the origin along the normal vector w. Choose w and b to maximize the margin, or distance between the parallel hyper planes that are as far apart as possible while still separating the data. These hyper planes can be described by the equationsW.X-b=1 and W.X-b=-1[17][18][19][20][21][22] to reduce the ambiguity prevalent when the data is not classified properly. The results section deals with regression specifically R square more on which is explained in[26].

4. Results

The proposed research work for SVM would provide more appropriate results to the above existing problem by performing regression on group shopping explained in [25]which gives relationship between variables as shown below. Figure 6 Assumes 6 years in of advertising for online social group shopping to predict the sales accordingly as shown in table.

Years	Advertising X	Sales Y
1	10	15
2	20	28
3	30	42
4	40	56
5	50	70
6	20	19

Figure 6: Advertising vs.S

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.985367
R Square	0.970948
Adjusted R Square	0.963685
Standard Error	4.134284
Observations	6

PROBABILITY OUTPUT

Percentile	Sales
8.333333	15
25	19
41.66667	28
58.33333	42
75	56
91.66667	70

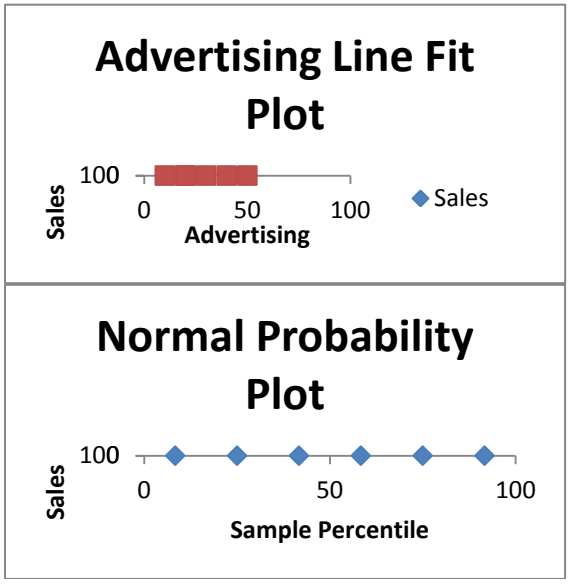

Figure 7: Plots for Advertising and Normal probability plot

As R square is above 70%i.e.0.970948, in regression statistics, there exists avery strong relationship between the variables X and Y.

5. CONCLUSION AND FUTURE SCOPE

The paper discuses Different IR models, from these IR models it can be concluded that the BM25 model, f(q,d) is represented by a conditional probability distribution P(r|q,d) where r takes the value of 1 means relevant or 0 means non relevant. The Language Model for IR (LMIR), f (q, d) is represented as a conditional probability distribution P (q|d). The probability models can be calculated with the words appearing in the query and document, and thus no training is needed .InLearning to rank given a query the document is ranked and retrieved,where ranking model f(q,d) is used to sort the documents, where q denotes a query and d denotes a document.Machine learning is a technique to construct automatedranking model f (q, d) to retrieve the documents in search.The paper shows that for better classification the proposed method is Support Vector Machines which uses hyper planes which is a good technique for training data, which classifies huge amounts of high dimensional data into different classes.Since the data set considered her would be small future scope is to increase the data set and checkfor better performance.

6. REFERENCES

[1] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, 2009,Introduction to Information Retrieval, Cambridge University Press.

[2] M.Mitra, B.B. Chaudhuri,2000,Information Retrieval from Documents: A Survey,ACM,May 2000, Volume 2, Issue 2-3, pp 141-163

[3] Baeza-Yates, R. and Ribeiro-Neto, B. (2011).,Modern Information Retrieval - the concepts and technology behind search.Addison Wesley.

[4] Croft, Metzler, and Strohman, 2010,"Search Engines: Information Retrieval in Practice," According to Information Retrieval in Practice by Croft, Metzler, and Strohman

[5] Djoerd Hiemstra,University of Twente ,Goker, A., and Davies, J.,November 2009, Information Retrieval: Searching in the 21st,Century. John Wiley and Sons, Ltd., ISBN-13: 978-0470027622

[6] C. Zhai and J. Lafferty., 2004,A Study of Smoothing Methods for Language Models Applied to Information Retrieval. ACM Transactions on Information Systems (TOIS): 22(2)

[7] SALTON, G., and C. BUCKLEY ,1990. "Improving Retrieval Performance by Relevance Feedback." Journal of the American Society for Information Science, 41(4), 288-97.

[8] H. R. Turtleand,W.B. Croft.,1991,Evaluationof aninferencenetwork-basedretrievalmodel. ACM Trans.Inf. Syst.,9(3):187–222,July1991

[9] Metzler, D. and Croft, W.B.,2004 "Combining the Language Model and Inference Network Approaches to Retrieval," Information Processing and Management Special Issue on Bayesian Networks and Information Retrieval, 40(5), 735-750, 2004

[10] Metzler, D., and Manmatha, R., ,2004"An Inference Network Approach to Image Retrieval," Proceedings of the International Conference on Image and Video Retrieval (CIVR 2004), 42-50, 2004

[11] Tie-Yan Liu (2009) "Learning to Rank for Information Retrieval", *Foundations and Trends® in Information Retrieval*: Vol. 3: No 3, pp 225-331. http:/dx.doi.org/10.1561/1500000016

[12] Wang, L., Lin, J., and Metzler, D. ,2010,"Learning to Efficiently Rank," in the Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2010), 2010.

[13] Hang LI Oct 2011,A Short Introduction to Learning to Rank, Special Section on Information-Based Induction Sciences and Machine Learning, IEICE TRANS. INF. & SYST., VOL.E94–D, NO.10

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

[15] V. Bush, 1945,'As we may think', The Atlantic Monthly, vol. 176, no. 1, pp. 101-10,.

[16] C. N. Mooers,1950, 'The theory of digital handling of non-numerical information and its implications to machine economics', in Association for Computing Machinery Conference, Rutger University.

[17] C. Cortes, V. Vapnik, Support Vector Networks in Machine Learning ,1995,20, pp. 273-297

[18] Christopher J.C. Burges, 1998, "A Tutorial on Support Vector Machines for Pattern Recognition", Data Mining and Knowledge Discovery, 2, 121-167, Kluwer Academic Publishers

[19] Tristan Fletcher, March 1, 2009 "Support Vector Machines Explained ", ucl

[20] Dustin Boswell, August 6-2002, Introduction to Support Vector Machines,

[21] J.P. Lewis, Dec-2004, A Short Support Vector Machine Tutorial,CGIT Lab/IMSC,U.South California

[22] John C. Platt,1999, Fast Training of Support Vector Machines using Sequential Minimal Optimization,Advances in kernel methods ,Pages 185 - 208 ,MIT Press Cambridge, MA, USA ©1999,table of contents ISBN:0-262-19416-3

[23] M. IkonomakisS, S. Kotsiantis, V. Tampakas, August 2005, Text Classification Using Machine Learning Techniques, WSEAS Transactions on Computers, Issue 8, Volume 4, pp. 966-974

[24] James G.Shanahan,Norbert ,2003,Improving SVM Text Classification Performance through Threshold Adjustment ,ML:ECML 2003,Volume 2837,pp 361-372

[25] Prasanna Kothalkar,Priyanka Desai,March ,2012, mplementing Social Group Shopping using Support Vector Machines, IJCA Journal, icwet2012 - Number 1IJCA Proceedings on International Conference and workshop on Emerging Trends in Technology (ICWET 2012) icwet(1):35-40,

[26] B. G. Kermani, I. Kozlov, P. Melnyk, C. Zhao, J. Hachmann, D. Barker, and M. Lebl , 2007 ,Using Support Vector Machine Regression to Model the Retention of Peptides in Immobilized Metal-affinity Chromatography, ns Actuators B Chem. 2007 July 16; 125(1): 149–157.