Asian Journal of Information Technology 5 (1) : 79-85, 2006

Web Document Representation Based on Knowledge Granularity

[1]Faliang Huang, [2]Shichao Zhang
[1]Department of Computer Science, Fujian Normal University, Fuzhou, China
[2]Department of Computer Science, Guangxi Normal University, Guilin, China

Abstract: Clustering Web documents is a fundamental task in Web Mining. Clustering analysis assists in reducing search space and decreasing information retrieval time. In this study we present a new data model for Web document representation based on granulation computing, named as Expanded Vector Space Model (EVSM), which facilitates knowledge engineers to acquire and understand processing results. We experimentally evaluate the proposed approach and demonstrate that our algorithm is promising and efficient.

Key words: Web document, knowledge, granularity

INTRODUCTION

In an effort to keep up with the tremendous growth of the World Wide Web (WWW), many research projects were targeted on how to organize such information in a way that makes it easier for the end users to find the needed information efficiently and accurately. Information on the web is mainly presented in the form of text documents (formatted in HTML). Clustering analysis is an important way of organizing information. It assists in reducing search space and decreasing information retrieving time. It is also useful for improving the recall and precision of IR systems and personalizing search engines effectively. Based on knowledge granularity theory[1-5] and article structure principle[6], in this study we present a novel scalable clustering algorithm for web document clustering.

This study is organized as follows. Section 2 describes the process of clustering web document clustering. Section 3 proposes a model to represent web documents. Section 4 designs an algorithm for clustering web documents. In section 4, several experiments are conducted for evaluating the proposed approach. In the last section we conclude this paper.

DESCRIPTION OF WEB DOCUMENT CLUSTERING PROBLEM

Web document clustering is rooted in text data mining techniques and share many concepts with traditional data clustering methods. Generally speaking, web document clustering methods attempt to segregate the documents into groups where each group represents a certain topic that is different from those topics represented by other groups. Currently there are two types of web document clustering in general: online web document clustering and offline web document clustering. The work in this study is focused on the second one.

Extant methods used for text clustering include decision trees[7-20], statistical analysis[7] and neural nets[8,9,21]. These methods are at the cross-roads of more than one research area, such as Database (DB), Information Retrieval (IR), and Artificial Intelligence (AI) including Machine Learning (ML) and Natural Language Processing (NLP). The existing techniques for clustering web documents rely on the following steps:

- Based on a given data representation model, a web document is represented as a logic data structure.
- Similarity between documents is measured by using some similarity measures that is depended on the above logic structure.
- With a cluster model, a clustering algorithm will build the clusters using the data model and the similarity measure.

Most of web document clustering methods that are in use today are based on the Vector Space Model, which is a very widely used data model for text classification and clustering. The VSM represents a web document as a feature vector of the terms that appear in that document. Each feature vector contains term weights (usually term-frequencies) of the terms appearing in that document. Similarity between web documents is measured by distance of the corresponding vectors. In Vector Space Model, the cosine measure and the Jaccard measure are the most common similarity measures. The aim of computing weight of a selected term is to quantify the term's contribution to ability to represent the source document topic. The focus of the Vector Space Model is how to choose terms from documents and how to weigh the selected terms.

Corresponding Author: Faliang Huang, Department of Computer Science, Fujian Normal University, Fuzhou, China

Choosing terms from a document: In essence, choosing terms from documents is actually a feature selection problem. In web document preprocessing the following parsing and extraction steps are needed:

- Ignoring case, extract all unique terms from the entire set of documents.
- Eliminate non-content-bearing stopwords such as a and, the, etc.
- For each document, count the number of occurrences of each term.
- Using heuristic or information-theoretic criteria, eliminate non-content-bearing high-frequency and low-frequency terms.
- After the above elimination, one term of the remaining terms is considered as one feature of the web document.

In this process, the Step 3, i.e. how to filter out the so-called useless terms or how to define the concept uselessness, is a headachy problem.

Weigh selected terms: In VSM, term weights are calculated based on the following two factors: term frequency, f_{ij}, the number of occurrence of term y $_j$in document x_i and inverse document frequency, log(N/dj), where N is the total number of documents in the collection and d_j is the number of documents containing term y_j.

The similarity $sim(x_i, x_j)$, between one document x_i and another document x_j, can be defined as the inner product of document vector X_i and document vector X_j:

$$sim(x_i, x_j) = X_i \cdot X_j = \frac{\sum_{k=1}^{m} w_{ik} \cdot w_{jk}}{\sqrt{\sum_{k=1}^{m} w_{ik}^2 \cdot \sum_{k=1}^{m} w_{jk}^2}} \quad (1)$$

here m is the number of unique terms in the document collection. Weight w_{ik} of document x_i is ▶. Apparently, the larger number of the same terms and the greater weight of the ones contribute to the greater similarity between documents.

MOTIVATIONS

Granularity theory: Granulation computing is a natural problem-solving methodology deeply rooted in human thinking; it is intrinsically fuzzy, vague and imprecise. Researchers have idealized it into the notion of partition, and developed into a fundamental problem solving methodology. Pawlak[13,14] supposed that man's intelligence is just the ability to classify. When investigating fuzzy information Zadeh[4,5] define three important concepts: granulation, organization and causation, on this base he considers granulation as a large umbrella which envelops all researches concerning granulation theories, methodologies, techniques and tools. Y.Y.Yao[1-3] and his collaborators conduct deeper study and present to solve consistent classification problem with lattices composed of every partition. Those works provide new methods and thinking ways.

As to essence, knowledge granularity is data set characterized as similarity in reasoning. Knowledge granularity with sufficiently conceptual sentences is beneficial for knowledge engineers to understand valuable relations hidden in data repository. With granularity calculation data can be more efficiently and effectively disposed of and knowledge engineers can handle the same dataset in different lays, this provides more reliable soundness for interpreting results of various data analysis methods. Virtually the procedure to construct knowledge granularity is the process to preprocess and convert data to be managed. In another word, granular computing is just one sub-problem of knowledge representation domain. Presently mainstream forms of granular computing are as follows:

- **Fuzzy set:** Fuzzy set, introduced by Zadeh in 1965, is a generalization of classical set theory that represents vagueness or uncertainty in linguistic terms. In a classical set, an element of the universe belongs to, or does not belong to, the set, i.e., the membership of an element is crisp—either yes or no. A fuzzy set allows the degree of membership for each element to range over the unit interval.
- **Rough set:** The key strength of rough set theory (partition) is the capability of processing knowledge in terms of approximation by partitions, table representations and quotient sets (knowledge level information). For general granulation such capability is not available yet. The knowledge processing can be expressed by approximation of granules, table representation and quotient sets (knowledge level processing) in the setting of pre-topological spaces.

Article structure principle: According to article structure theory[6], article structure is composition of article content, which is dialectic unity between intrinsic orderliness and law of objective things and author's subjective cognitions (observations, imaginations, etc) of the objects. Article structure plays a very important role in quality of an article. Paragraph is a smallest and comparatively independent unit to construct an article and is usually used to express the author's viewpoint. Readers often are only interested in some paragraphs of a lengthy article in the course of reading. Consequently, paragraph is a significant logical layer of representing a web document and is an important granularity.

Intrinsic limitations of VSM: It is well-known that web document representation model is of importance to quality of web document clustering results. VSM is a common and successful data model for web document, but after analysis, it is not difficult to discover there are some limitations in it.

Interoperability of OLAP operations: In VSM, suppose we treat a term as a feature of a web document object, the document collection can be viewed as a multi-dimensional database. Traditional data mining techniques reveal that such Online Analytical Processing (OLAP) operations as roll-up and drill-down can facilitate knowledge engineers acquiring and understanding information in multi-level granularities[10,19], however, traditional VSM provides only two level granularities, that is to say, document-term, the span between document level granularity and term level granularity is too far to make the previous OLAP operations fail to work. On this base, a web document can be represented as another logic model by adding a new granularity.

Document false correlation: The paragraph level granularity is excluded by traditional document representation model characterized as document-term two-level granularity. Owing to the exclusion, knowledge engineers are frequently confronted with the document false correlation, depicted as (Fig. 1), in the course of clustering web documents.

Example 1. Let d1 = {p1, p2, p3},d2 = {p1, p2} be two documents, d1 and d2 are represented the same term(feature) vector composed of term t1 and t2 after preprocessing. The conclusion that document d1 and d2 are very similar or even identical can be drawn from comparing the two documents at the document granularity level. However, performing paragraph granularity level comparison between above two documents probably results in conclusion that there is some differences among them. What on earth leads to the inconsistent conclusions? Dipping into distribution of terms will reveal the hidden truth: global distribution of term t1 and t2 is the same but local distribution of them is different.

Frequent Occurrence of zero-valued Similarity: As we have seen, in VSM a single document is usually represented by relatively few terms. The document vector which is characteristic of high-dimension and sparseness results in zero-valued similarity which deceases quality of clustering when define the relation between document and document.

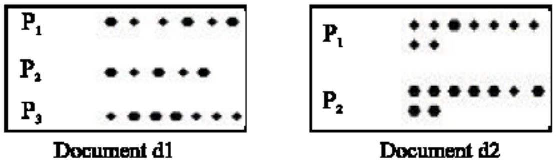

Fig.1: Document false correlation

DATA MODEL

Concerned with advantages and limitations of traditional VSM, we present an Expanded Vector Space Vector (EVSM) model in which web document is represented as a Document-Paragraph-Term (D-P-T) configuration characterized as multi-level and multi-granularity and paragraph granularity is computed with the guidance of tolerance rough set theory.

D-P-T Configuration: In this framework, a web document is represented as following logic layers:

- Document layer □

D = {DId, Title, Body, Length}
Body = {P_1, P_2, …, P_n}

Here D is a web document, Did is id of the web document, Title is title of the web document, body is body of the web document which is composed of a paragraph set, length is total length of the paragraphs.

- Paragraph layer:

P = {PId, DId, Position, Length, Term, TRRate}
Term = {term_1, term_2, …, term_n}

Here P is a paragraph of a web document, PId is id of the paragraph, DId is the id of the web document containing current paragraph, Position is position of the paragraph which falls into three classification: Front, Middle and End. Length is length of the paragraph, Term is a term set of the paragraph, TRRate denotes term repeating rate in the paragraph.

- Term layer □

term = {TId, PID, Position, Weight}

Here term is a term of a paragraph, Tid is id of the term, Pid denotes the id of the paragraph containing current term, Position denotes attribute of html tag enclosing the current term. Weight denotes a weight produced from a weighing system.

EVSM based on Tolerance Rough Set: Tolerance Rough Set Model (TRSM) is an expanded model of the classical rough set model[13,14]. In this model a tolerance relation T, upper approximation $B^{-}(X)$ and lower approximation $B^{-}(X)$ are defined as below:

$$T=\{(x,y)|x\in U\wedge y\in U\wedge\forall c_i(c_i\in B\Rightarrow(c_i(x)=c_i(y)\vee c_i(x)=*\vee c_i(y)=*))\} \tag{2}$$

$$B_{-}(X)=\{x\in U\,|\,I_B(x)\subseteq X\} \tag{3}$$

$$B^{-}(X)=\{x\in U\,|\,I_B(x)\cap X\neq\phi\} \tag{4}$$

With above TRSM, we apply granular computing to paragraph level granularity. For a paragraph we can define an indiscernible relation I, tolerance relation Ψ, upper approximation $\psi^{-}(X)$ and lower approximation $\psi^{-}(X)$ as following:

$$I_\lambda(t_i)=\{t_j\,|\,f_P(t_i,t_j)\geq\lambda\}\cup\{t_i\} \tag{5}$$

$$t_i\Psi t_j\Leftrightarrow t_j\in I_\lambda(t_i) \tag{6}$$

$$\psi_{-}(X)=\{t_i\in T\,|\,\frac{|I_\lambda(t_i)\cap X|}{|I_\lambda(t_i)|}=1\} \tag{7}$$

$$\psi^{-}(X)=\{t_i\in T\,|\,\frac{|I_\lambda(t_i)\cap X|}{|I_\lambda(t_i)|}>0\} \tag{8}$$

Suppose X is a term set expressing a vague concept, $\psi^{-}(X)$ is core connotation of the concept and $\psi^{-}(X)$ is extension of the concept. Occurrence frequency of zero-valued similarity can be greatly lessened by using upper approximation of the concept expressed by paragraph level granularity knowledge.

Improved TFIDF Weighing System in EVSM Model: We produce an improved TFIDF weighing system based on the traditional TFIDF weighing system of VSM[10,17]. $p_i=\{t_1,t_2,\ldots,t_n\}$ is a paragraph of a web document and its upper approximation is $p_i=\{t_1,t_2,\ldots,t_m\}$,$w_{ij}$ denotes weight of term t_j in paragraph p'_i□w'_{ij} is normalized value of weight w_{ij}, both weight are formalized as below:

$$w_{ij}=\begin{cases}1+f_{p_i}(t_j)\times\log\frac{N}{f_P(t_j)} & \text{if } t_j\in p_i\\ (\min_{t_k\in p_i}w_{ik})\times\frac{\log\frac{N}{f_P(t_j)}}{1+\log\frac{N}{f_P(t_j)}} & \text{if } t_j\in p'_i\wedge t_j\notin p_i\\ 0 & \text{if } t_j\notin p'_i\end{cases} \tag{9}$$

$$g\,w'_{ij}=\frac{w_{ij}}{\sum_{t_j\in p'_i}w_{ij}} \tag{10}$$

To demonstrate the use of the EVSM framework, we detail the process of converting web document by an example as follows.

Example 2. Let paragraph collection be P = {p_1, p_2, p_3, p_4, p_5, p_6, p_7}, term collection be T = {t_1, t_2, t_3, t_4}, the frequency data is listed in Table 1.

Let threshold| λequals 4, with Eq. 3,4 and 5 upper approximations of the paragraph pi (I = 1, 2 , ..., 7) can be computed as below:

$$\psi^{-}(p_1)=\psi^{-}(p_2)=\psi^{-}(p_4)=\{t_1,t_2,t_3,t_4,t_5\}$$

$$\psi^{-}(p_3)=\psi^{-}(p_5)=\psi^{-}(p_6)=\{t_1,t_2,t_4,t_5\}$$

$$\psi^{-}(p_7)=\{t_3,t_4,t_5\}$$

We weigh the paragraph p_1 with traditional TFIDF and p'_1 with the improved TFIDF, result is listed in Table 2.

Evaluation on paragraph granularity's representing ability: In order to label document according to paragraph clustering results, it is necessary to develop appropriate metrics to evaluate paragraph granularity's ability to represent its parent web document's topic. For measuring the representative ability, we here extract three important attributes from each paragraph: Paragraph Position, Term Repeating Rate and Paragraph Relative Length.

Paragraph position: We classify all paragraphs in one web document into by paragraph position in web document: Type of the first paragraph is Front, type of the last paragraph is End, and type of other paragraphs is Middle. On this base, we present a strategy to determine weight of the position attribute of the paragraph.

Let p_i be a paragraph of web document d, |d| denotes the total of paragraphs of a web document. ▶ denotes position weight of the paragraph p_i.

$$p_i.PP=\begin{cases}1 & \text{if } |d|=1\\ 1/2 & \text{if } |d|=2\\ 1/3 & \text{if } |d|\geq 3 \text{ and } (i=1 \text{ or } i=|d|)\\ \frac{1}{3*(|d|-2)} & \text{if } |d|\geq 3 \text{ and } 2\geq i\geq |d|-1\end{cases}$$

Table 1: Sample Paragraph-Term Frequency Array

Term/paragraph	p_1	p_2	p_3	p_4	p_5	p_6	p_7
t_1	0	8	3	6	7	1	0
t_2	0	0	7	5	2	6	0
t_3	5	3	0	4	0	0	1
t_4	2	0	4	0	5	6	4
t_5	3	7	5	2	5	4	0

Table 2: VSM and EVSM

	Improved TFIDF			Traditional TFIDF	
	weight			weight	
term			term		
	Non-normalization	normalization		Non-normalization	normalization
t_1	0.093	0.015	t_1	0	0
t_2	0.143	0.023	t_2	0	0
t_3	1.731	0.281	t_3	0.731	0.25
t_4	2.089	0.339	t_4	1.089	0.37
t_5	2.104	0.342	t_5	1.104	0.38

Paragraph relative length: According to article structure theory, generally speaking, the more detailed a paragraph description is, the more important a paragraph is to the parent web document. So we give the following definition (Paragraph Relative Length, abbreviated as PRL):

$$p_i.PRL = \frac{p_i.Length}{d.Legnth} \tag{11}$$

Term repeating rate: Article structure principle holds that high some terms occur very frequently in some position to give importance to some viewpoint of author. We define Term Repeating Rate (TRRate) as the following formula:

$$p_i.TRRate = \frac{\sum_{term_j \in p_i.Term} freq(term_j)}{p_i.Length} \tag{12}$$

From above three measures, we can define weight w_{pi} of the paragraph as below:

$$w_{p_i} = PW * p_i.PP + LW * p_i.PRL + TRW * p_i.TRRate$$
$$\text{s. t. } PW + LW + TRW = 1$$

here PW,LW and TRW respectively denotes contribution to the paragraph representative ability of the attribute Paragraph Position, Paragraph Relative Length and Term Repeating Rate. The concrete values of PW, LW and TRW can be manually given by domain experts or automatically given by computer.

ALGORITHM DESIGN

Labeldocument algorithm: For simplicity the main procedure is described as following: first, score each paragraph by attribute Paragraph Position, Paragraph Relative Length and Term Repeating Rate. Second, assign document to the optimum by the value of membership to topic cluster, motivated by high-voting principle of multi-database mining[18].

```
Algorithm□ LabelDocument
Input □ web document d= (title, $p_1,p_2$ ,..., $p_n$),topic set
T = ($T_1,T_2,..., T_n$)
Output□ label of web document d
Method□
(1) for each $p_i \in$ d do
compute $w_{pi}$□
end for
(2) for each $T_j \in T$ do
if title $\in$ d $w_j$ = TW
        for each $p_i \in$ d do
if $p_i \in T_j$ then $w_j = w_j$ + PSW*$w_{pi}$
  end for
end for
(3)      $label = \arg\max_{T_j \in T}(w_j)$
return label.
```

WDCBKG algorithm: Input Web document collection D number of clusters K term frequency threshold β tolerance threshold λ ,minimal change rate ε
Output: K web document clusters T_1, T_1 ,..., T_k

- preprocess web document collection and convert it paragraph vectors with the guidance of the data model EVSM.
- cluster paragraphs with k-means
- label the web documents with LabelDocument.

EXPERIMENTS

Dataset selection: To evaluate our proposed algorithm WDCBKG, we download 15013 web documents from sub-directory of Yahoo! News. The documents distribution is listed in Table 3.

Experimental results: n this section, we evaluate the function of the approach. The following experiments were conducted on a Dell Workstation PWS650 with 2 GB main memory and Win2000 OS.

We access our proposed approach from three aspects as following:

Performance of clustering results: We use F-measure, which is the harmonic mean of values of precision and recall rate, to evaluate clustering results by comparing WDCBKG algorithm with VSM_Kmeans algorithm. We randomly select 10 groups of web documents from the document collection and cluster each group data, the size

Table 3: Distribution of web document collection

Group NO	Label	Number of Web Document
1	Sports	2566
2	Health	2641
3	Technology	2309
4	Business	2470
5	Politics	2163
6	Label	2566

Table 4: The comparison of clustering results of WDCBKG and VSM_Kmeans

Group NO	VSM_Kmeans	WDCBKG
1	0.616	0.768
2	0.592	0.744
3	0.626	0.783
4	0.607	0.776
5	0.621	0.765
6	0.631	0.756
7	0.612	0.78
8	0.584	0.74
9	0.625	0.772
10	0.598	0.771

of which is 10000, with VSM_Kmeans and WDCBKG respectively. Table 2 shows the results of the two algorithms. From table 2 we can see that, compared to VSM_Kmeans, performance of WDCBKG is great improved.

Scalability: We conduct a group of experiments with different data set that is of different size. From Figure 2 we can see that the performance of clustering results from EVSM_WDCBKG doesn't decease with the size of experimental data set increased but keep satisfied stability, fluctuating from 0.75 to 0.81. Consequently, as far as data set size is concerned, our approach is scalable.

Sensitiveness to tolerance threshold parameter: Tolerance threshold parameter is rather important to our WDCBKG. From our EVSM model it is not difficult to get such deduction that inadequate tolerance threshold can decrease the performance of the clustering results: on one hand, too small tolerance threshold can add noise data while representing clustering objects, on the other hand, too large tolerance threshold can make EVSM tend to VSM, both cases can lead to worse performance. From Fig. 3 we can understand our experimental result corresponds to our deduction: when tolerance threshold equals 5, the performance is the best, however, when it equals 2,3 or 8, the performance is worst.

CONCLUSIONS

The rapid development and prevalent use of the Web highlight the need for new technologies for design, implementation and management of Web-based information systems. Web document clustering has become an important research area in web mining. Vector Space Model plays an important role in the researches covering a broad spectrum of topics. Though the model is constructed on a stable probability theory and simplifies computation, there are still some limitations in it.

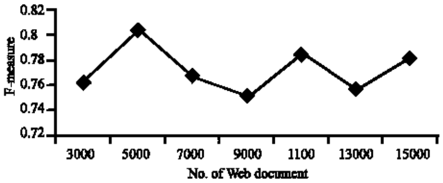

Fig.2. Scalability of WDCBKG

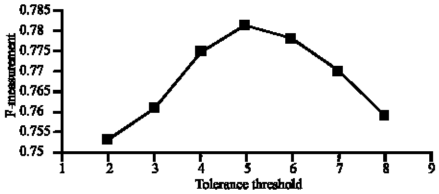

Fig.3: Sensitiveness to tolerance threshold

In this study we have studied the intrinsic limitations of Vector Space Model and proposed a new representation model, named as EVSM model. To evaluate our approach, we have conducted some experiments. The results have shown that our algorithm is effective, efficient and promising.

REFERENCES

1. Yao, Y.Y., 2001. Information granulation and rough set approximation, Intl. J. Intelligent Sys., 16: 87-104.
2. Yao, Y.Y., 2003. Granular computing for the design of information retrieval support systems, in: Information Retrieval and Clustering, Wu, W., Xiong, H. and S. Shekhar, (Eds.), Kluwer Academic Publishers 299.
3. Yao, Y.Y., 2004. A Partition Model of Granular Computing. T. Rough Sets., pp : 232-253.
4. Zadeh, L.A., 1997. Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, Fuzzy Sets and Systems, 19: 111-127.
5. Zadeh, L.A., 1998. Some reflections on soft computing, granular computing and their roles in the conception, design and utilization of information/ intelligent systems, Soft Computing, 2: 23-2.
6. Zheng Wenzhen, 1984. Architecture for Paragraphs (in Chinese). Fujian People's Press.
7. Bing, Liu, X. Yiyuan, S. Yu. Philip, 2000. Clustering Through Decision Tree Construction In SIGMOD-00.
8. Hung, C. and S. Wermter, 2003. A dynamic adaptive self-organising hybrid model for text clustering, Proceedings of The 3rd IEEE International Conference on Data Mining (ICDM'03), Melbourne, USA, pp: 75-82.

9. Hung, C. and S. Wermter, 2004. A time-based self-organising model for document clustering, Proceedings of International Joint Conference on Neural Networks, Budapest, Hungary,pp: 17-22.
10. Chi Lang Ngo, Hung Son Nguyen, 2004. A Tolerance Rough Set Approach to Clustering Web Search Results. PKDD pp: 515-517.
11. Han, J. and M. Kamber, 2000. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers.
12. Yoon, J., V. Raghavan and Venu Chakilam, 2001. BitCube: Clustering and Statistical Analysis for XML Documents. 13th International Conference on Scientific and Statistical Database Management, Fairfax, Virginia pp:18-20.
13. Kryszkiewicz, M., 1998. Properties of in complete information systems in the framework of rough sets. In:L Polkowski, A Skow roneds. Rough Sets in Data Mining and Knowledge Discovery. Berlin: Springer-Verlag, pp: 422-450.
14. Kryszkiewicz, M., 1998. Rough set approach to incomplete information system. Information Sciences, 112: 39-495.
15. Pawlak, Z., 1991. A Rough Sets, Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Dordrecht.
16. Pawlak, Z., Granularity of knowledge, indiscernibility and rough sets, Proceedings of 1998 IEEE International Conference on Fuzzy Systems, 106-110.
17. Salton, G. and J.M. McGill, 1983. (Eds.): Introduction to Modern Information Retrieval, McGill-Hill.
18. Zhang, S., 2001. Knowledge discovery in multi-databases by analyzing local instances. PhD Thesis, Deakin University, 2001.
19. Viette Poe, Patricia Klauer and S. Brobst, 1997. Building A Data Warehouse for Decision Support. Prentice Hall PTR; 2nd (Edn.).
20. Yang, Y. and J.O. Pedersen, 1997. A comparative study on feature selection in text categorization.In Proceedings of the Fourteenth International Conference on Machine Learning. San Francisco: Morgan Kaufmann, pp: 412–420.
21. Hsu, A.L. and S.K. Halgamuge, 2003. Enhancement of topology preservation and hierarchical dynamic self-organising maps for data visualization, Intl. J. Approximate Reasoning, 32: 259-279.