ISSN : 0974 - 7435 Volume 8 Issue 6

BioTechnology
An Indian Journal

FULL PAPER

BTAIJ, 8(6), 2013 [787-794]

An chinese text classification algorithm based on graph space model

Xiaoqiang Jia
School of Mathematics and Information Science, Institute of Applied Mathematics, Weinan Normal University, Weinan 714000, Shaanxi, (CHINA)
E-mail: 394892738@qq.com

ABSTRACT

In the field of information processing, most of the existing text classification algorithm is based on vector space model, but vector space model is not able to effectively express the document structure information, so that it is not enough to express the semantic information of documents context. In order to get more semantic information effectively, by the study of text representation of graph space model, use Common node structural equivalence and Common chain structure equivalence, analyse nodes and edges of the maximum common substructure graph, and judge which if is a true semantic equivalence. Next, a data structure for text classification on Graph space model was designed. On the basis of structural equivalence analysis, the distance formula of "MCS" has been improved, then an improved text similarity metric algorithm based on the graph space model has been proposed, experiments show that the text classification method is effective and feasible.

KEYWORDS

Structural equivalence;
Text classification;
Graph space model;
Maximum common subgraph;
Similarity.

INTRODUCTION

Large text is often hidden valuable information. In the real world, the most information need to be dealt with can be transformed into text storage and representation. The text is the most important information carrier. Text classification is given in the set of labels in advance, according to the text content determine text classification. As an important branch of data mining, text classification has been widely applied in the field of information filtering, information retrieval, natural language processing.

Research on text classification began in the nineteen fifties, Luhn[1] in this field made groundbreaking research, he first lead into statistical thoughts for text data classification. In 1960 Maron[2] in Journal of ACM published the first paper on text data classification. In 1963, Borko et al proposed a classification analysis method for document utilization factor. Subsequently many scholars in this field made very fruitful research. Research on text classification can be divided into three stages: the first stage is the nineteen eighties, this stage mainly research on classification theory; the second stage is the nineteen eighties to nineteen nineties, experimental study on automatic classification; the third stage is nineteen nineties, later, practical stage with the development of automatic classification. The Internet technology makes the text data increased dramatically, this time-consuming, poor flexibility, use difficult increasingly and unable to meet the needs of practical application, and can be gradually replaced by methods of machine learning[3].

Vector space model (Vector Space Model) was put forward by Professor Salton et al in 1968 and developed the document representation method. However, because the vector space model is a model does not consider the feature order of bag of words of text, though this model brought in calculation and the convenience of operation, but lost a lot of information of text structure, as well as lack the information of feature term context. The text structure information or context in natural language is essential. Therefore, from the perspective of natural language, vector space model is still not perfect. Aim at vector space representation model defects, many scholars proposed document representation method based on the graph model. As presented by Svetlana in his paper[4] on Verb Net and the document representation model based on concept graph; Bhoo pesh and Pushpak in their paper[5] proposed to construct feature vector to represent a document according to the UNL[6], and the text clustering used by SOM technology; and Inder Jeet and Eric in their paper[7] also presented a representation for a document model for multi-document summarization extraction. The graph model although well reflects the semantic information of document, but is too complex to give measure similarity criteria, and some also need additional supporting information. Recently, Adam Schenker etal in their paper[8] proposed a simple method based on graph model of the document, but their model is based mainly on the position of boolean association characteristics, and did not consider the frequency of feature items appears. So their model has to be further modified and perfected. Kadd etc proposed based on word co-occurrence relationship between adjacent patterned text representation method, which is suitable for the independent significance of the word in English, is not suitable for Chinese character, and also ignore the weight information of edge.

As a result of the vector space model can not effectively express the structure of the text information, a text representation method based on the graph space model is researched. And on the basis of the structural equivalence, further analysis of the maximum common substructure graph nodes and edges are “real” semantic equivalence, an improved text similarity standard is proposed. And applied it to text categorization, the experimental results shows text classification method based on the graph space model is feasible and effective.

definitions relate to Graph model

Definition 1: Graph is a data structure $G=(V,E,\alpha,\beta)$, in which

$V(G)$ is nodes finite nonempty set in G.

$E(G)\subseteq V\times V$ is all the edges set in G.

$\alpha: V\to\Sigma_V$ is node marked function in G.

$\beta: E\to\Sigma_E$ is edge marked function in G.

Σ_V and Σ_E mean nodes and edges labels set in G, under simple case, a node or edge is only one label, in the graph space model of text, the general node label information represent the node term, and edge label information reflect if the two node is adjacent. Sometimes in order to distinguish effectively different nodes and edges, nodes or edges also can have multiple labels, such as node label can be term frequency, document frequency and the location information of term appearing in the text, edge label is the number two adjacent nodes appearing in text or text set and text position information. Given a graph data structure in definition 1, nodes and edges are only one label, and if there is no special, node labeled terms, edges marked if the two node were adjacency information.

Definition 2: G_1 is the subgraph of G_2, it is marked with $G_1\subseteq G_2$, if $V_1\subseteq V_2$ $E_1\subseteq E_2\cap(V_1\times V_1)$, $\forall x\in V_1$, $\alpha_1(x)=\alpha_2(x)$ was established. $\forall e=(x,y)\in E_1$,there was $\beta_1(e)=\beta_2(e)$. On the contrary, G_2 is the supergraph of G_1.

Definition 3: G_1 and G_2 are for graph isomorphism, simple written as $G_1\cong G_2$, if there exists a bijective function $f:V_1\to V_2$, there is $\forall x\in V_1$, $\alpha_1(x)=\alpha_2(f(x))$ holds. $\forall e=(x,y)\in E_1$, there exists $e'=(f(x),f(y))\in E_2$, $\beta_1(e)=\beta_2(e')$, holds, and $\forall e'=(x',y')\in E_2$, there exists $e=(f^{-1}(x'),f^{-1}(y'))\in E_1$, $\beta_1(e)=\beta_2(e')$ holds. If $V_1=V_2=\varnothing$, then G_1 and G_2 are called the null graph isomorphism. If there exists a bijective function $f:V_1\to V_2$, make G_1 and G_2 graph isomorphism[9-11], and G_2 is the subgraph of G_3, then the subgraph of G_1 is isomorphism with G_3.

Definition 4 : There exist graph G, G_1 and G_2, if the subgraph of G is isomorphism with G_1,and the subgraph of G is isomorphism with G_2, then G is called the common subgraph of G_1 and G_2.

Definition5: G, G_1 and G_2 is graph, G is the maximum common subgraph[13,14] of G_1 and G_2, it is marked with $mcs(G_1, G_2)$, if G is the common subgraph of G_1 and G_2, there not exists other common subgraph G', then $|G'| > |G|$ holds.

The distance measurement method between graphs

The distance is a metric method of two object similarity, it mainly introduces several graph distance measurement method. If there is no special, $|G|$ means the size of G, that is the sum of graph nodes and edges, max{A, B} is the maximum number operations between A and B.

(1)MMCS

Fernandez and Valiente proposed MMCS formular, which is based on the maximum common subgraph

$$d_{MMCS}(G,G') = |MCS(G,G')| - |mcs(G,G')| \qquad \textbf{(1)}$$

(2)MMCSN

$$d_{MMCSN}(G,G') = 1 - \left|\frac{mcs(G,G')}{MCS(G,G')}\right| \qquad \textbf{(2)}$$

MMCSN is a form the result of MMCS standardization for the interval [0, 1]

STRUCTURAL EQUIVALENCE

Common node structural equivalence

Let $G_1 = (V_1, E_1, \alpha_1, \beta_1)$ and $G_2 = (V_2, E_2, \alpha_2, \beta_2)$ for a pair of graphs, $A_{n\times n}$ and $B_{m\times m}$ is respectively the adjacency matrix of G_1 and G_2, n and m respectively is the number of nodes of G_1 and G_2, in the case of A, adjacency matrix is constructed as follows:

$$A_{i,j} = \begin{cases} 1 & (i,j) \in E_1 \\ 0 & (i,j) \notin E_1 \end{cases} \quad \forall i,j \in V_1 \qquad \textbf{(3)}$$

An isolation term generally do not reflect certain semantic information, semantic of terms only can be obtained by analysis of other associated terms. In the adjacency matrix, the rows and columns of the matrix reflect the node degree information, namely node representation terms and the adjacency condition of other terms. Because the spatial model is undirected graph, so the row and column of matrix reflects the information is consistent. Then select two rows vectors of A and B respectively form two vectors $\xi = (A_{i,0}, A_{i,1}, \cdots, A_{i,n})$ and $\eta = (B_{i',0}, B_{i',1}, \cdots B_{i',m})$, and the two vector by computing correlation coefficient to measure the similarity of degree of the node "i" in G_1 and the node "i'" in G_2 ,that is structural equivalence degree of the term expressed by the node"i"and the term expressed by the node "i'" structural equivalence degree. In practice, although different terms may have the same semantic, but more complicated to analyze, so only to those nodes who represent the same terminology structure equivalence analysis, that is common node of the maximum common subgraph meet the $\alpha_1(i) = \alpha_2(i')$.

Generally speaking, two term tag for G_1 and G_2 set Σ_{V1} and Σ_{V2}, there will be $\Sigma_{V1} \neq \Sigma_{V2}$, So the dimension of the vector ξ and η is different, In order to facilitate comparison of correlations, it is need to be expand between G_1 and G_2, The purpose is to make $G_1' = (V_1', E_1', \alpha_1', \beta_1')$ and $G_2' = (V_2', E_2', \alpha_2', \beta_2')$ have the same term tag set after extending. Specific extensions in the case of G_1, as long as the terms in the Σ_{V2} don't appears in Σ_{V1}, the term is as an isolated node (node degree 0) to add to G_1. The expansion of G_2 is similar[15-17] with G_1.

The G_1 and G_2 were extended to G_1' and G_2', according to G_1' and G_2' to construct adjacency matrix A' and B', let $V_1' = \{0,1,\cdots,k-1\}$, $V_2' = \{0',1',\cdots,(k-1)'\}$, $k \in Z$. For convenient operation, A' is defined as follows:

$$A'_{u,v} = \begin{cases} 1 & (u,v) \in E_1' \\ 0 & other \end{cases} \quad \forall u,v \in V_1' \qquad \textbf{(4)}$$

B' is defined as follows:

$$B'_{u,v} = \begin{cases} 1 & (\alpha_2'^{-1}(\alpha_1'(u)), \alpha_2'^{-1}(\alpha_1'(v))) \in E_2' \\ 0 & other \end{cases} \quad \forall u,v \in V_1' \qquad \textbf{(5)}$$

Let the maximum common subgraph of G_1 and G_2 is G_{mcs}=$(V_{mcs}, E_{mcs}, \alpha_{mcs}, \beta_{mcs})$, then, compute structure equivalent degree of node $i \in V_{mcs}$, which can be obtained by $\xi' = (A'_{i,0}, A'_{i,1}, \cdots, A'_{i,k\text{-}1})$ and $\eta' = (B'_{i,0}, B'_{i,1}, \cdots B'_{i,k\text{-}1})$, which is composed of A' and B'. Firstly, the mean value and variance for row vector of the adjacency matrix are as follows:

$$\mu_{\xi'} = \frac{1}{n}\sum_j A'_{i,j} \qquad \textbf{(6)}$$

$$\sigma_{\xi'}{}^{2}=\frac{1}{n}\sum_{j}(A'_{i,j}-\mu_{\xi'})^{2} \tag{7}$$

$$\mu_{\eta'}=\frac{1}{n}\sum_{j}B'_{i,j} \tag{8}$$

$$\sigma_{\eta'}{}^{2}=\frac{1}{n}\sum_{j}(B'_{i,j}-\mu_{\eta'})^{2} \tag{9}$$

Then, the correlation coefficient of ξ' and η' is as follows:

$$\chi(i)=\chi_{\xi',\eta'}=\frac{\frac{1}{n}\sum_{k}(A'_{i,k}-\mu_{\xi'})(B'_{i,k}-\mu_{\eta'})}{\sigma_{\xi'}\sigma_{\eta'}} \tag{10}$$

Need to consider two special cases in the resulting formulas (13), when the denominator is 0, here is with $\sigma_{\xi'}$ example:

The first case, if all the dimension of the ξ' is 1, and then $\sigma_{\xi'}$ is 0, in practice this case in the experiments nearly does not appear.

The second case, if all the dimension of the ξ' is 0, said node "i" is an isolated point. Because of experimental structure graph is an undirected graph, and in the establishment of the adjacent table, first filter text segment only having single terms, so before G_1 expansion, there are not the isolated point, only generate in the process of graph expansion, but this kind of nodes in G_1 does not exist, so there is no necessary to do the node similarity.

Moreover what needs to explain, because in the text graph model has certain particularity, the product for the number of representative terms of node i in G_1 adjacent terms and the number of representative terms of node i in G_2 adjacent terms are generally less than $|\Sigma_{V1}\cup\Sigma_{V2}|$, this situation makes the node i representative terms in G_1 and G_2 without common adjacent terms, $\chi(i)<0$, This can actually think of node i representative terms in the two graph does not have the correlation.

So far, on the basis of the common node structure equivalence analysis, MCS formula can be changed as follows form:

$$d(G_1,G_2)=1-\frac{\sum_{i\in V_{mcs}}\chi(i)}{\max\{|G_1|,|G_2|\}} \tag{11}$$

Common chain structure equivalence

The Formula (14) only consider the node structure equivalent degree, in text the phrase structure also contain rich semantic information, therefore, on the basis of the node structure equivalence analysis, the common chain[18] structure equivalence analysis is given.

Definition 8 spanning tree refers to a linking graph, passing the edges set and all nodes of graph constitute the minimum connected subgraph of graph of being called depth-first traversal or breadth-first traversal (DFS), namely a minimum spanning tree of a connected graph.

Definition 9 Generation forest refers to an unconnected graph, each nodes set of connected component and the edges traveled together form a plurality of spanning tree, these spanning tree of connected graph constitutes spanning forest of unconnected graph.

Definition 10 common chain refers to spanning tree two of the maximum common subgraph in which all nodes number are greater than 1. Let $L=\{l_1,l_2,\cdots,l_n\}$ be common chain of the maximum common subgraph set $len(l)$, be the edges number of common chain, V_l be all the nodes set of common chain l, E_l be all the edges set of common chain l, then structural equivalence formula of common chain L is:

$$\overline{\chi}_l=len(l)\prod_{i\in V_l}\chi(i) \tag{12}$$

After increasing common chain structure equivalent, the formula (11) is changed into:

$$d(G_1,G_2)=1-\frac{\sum_{i\in V_{mcs}}\chi(i)+\sum_{l\in L}\overline{\chi}_l}{\max\{|G_1|,|G_2|\}} \tag{13}$$

At this point, measuring the distance function between graph and graph through structure equivalent of the common node and common chain is basically established. so, similarity formula[19-21] between two graphs is as follows:

$$d(G_1,G_2)=1-\frac{\sum_{i\in V_{mcs}}\chi(i)+\sum_{l\in L}\overline{\chi}_l}{\max\{|G_1|,|G_2|\}} \tag{14}$$

THE IMPROVED GRAPH SPACE MODEL

Data structure

In fact, many text classification plays an important role in the Statistical information has not been taken into account by formula (14). Generally speaking, Graph space model of the more important statistical information also includes the occurrence frequency and term weight information between two adjacent terms; here term weight information is mainly divided into two kinds. One is the term weight according to the various feature weighting formula; another is the location information term appeared in the Web text. In order to reflect the statistical information so much the better, the data structure of the definition of 1 is amended as follows:

Definition 11 Graph space data structure is $G=(V,E,\alpha_1,\alpha_2,\beta_1)$, in which:

$V(G)$ is node finite nonempty set of G.

$E(G)\subseteq V\times V$ is the edges set of G.

$\alpha_1:V\to\Sigma_V$ is the function of node labeled terms and one to one correspondence between the node and term G.

$\alpha_2:V\to\Sigma_W$ is marked as weight function in node G.

$\beta_1:E\to\Sigma_E$ is a common frequency function marked adjacent nodes in G.

The improved algorithm

Here, the distinction of the definition 11 and the definition 6.1 is expressed as the adjacent node co-occurring number by β_1, rather than merely reflected whether node pair is adjacency. Another one of the biggest difference is increased the node weights marker function α_2, the weight value information are reflected by α_2, let the node weight sequence is $(w_1,w_2,\cdots,w_n)$, each weight proportion respectively is $(x_1,x_2,\cdots,x_n)$, in which, $\Sigma_{xi}=1$, then, the weights of the node j in the text is for:

$$\alpha_2(j)=x_1w_1(j)+x_2w_2(j)+\cdots+x_nw_n(j) \tag{15}$$

If the training set is web formatted text, in text terms location information is as a part of weight, web text can be transformed into a DOM tree to obtain the node content, therefore, mark function of α_2 will be converted to the DOM tree to obtain three kinds of location information and give a certain weight, if the term appears in the location of <TITLE> value or the content position of <META> keywords position properties, then the node weight is 5, if the weight appear in the other positions for the node weight is 2, If a term appear in multiple locations, such as the term " Sports " appear in both the <TITLE> tags, but also In the <BODY> label, then regardless of the number of term node " Sports " occur in each position, the weight value is 7 (5+2=7).

According to the improved graph data structure, From the training text set construct a complex network, and then through the FS algorithm for feature selection, extraction of the various categories of 1000 characteristics will constitute the feature dimension reduction class diagram G_i, therefore, text categorization can actually be seen as similarity comparison of process of said to be test text graph and each class diagram after dimensionality reduction, similarity degree is bigger with G_i, and the more likely belongs to the category.

In the FS algorithm[22], feature extraction sequence can be used as a parameter of feature weights because the training text sets is not HTML/XML formatted text, so the term node j contains only the sequence weight information t (j) of feature discovery, so the node weights marked function is $\alpha_2(j)=t(j)$.

Adding node weight information, formula (10) is changed into:

$$\chi'(i)=\chi(i)\alpha_2(i) \tag{16}$$

After adding neighbor node co-occurring times and node weight information, formula (6.15) is amended as follows:

$$\overline{\chi}_l{}'=(lenl)(\sum_{e\in El}\beta_1')\prod_{i\in V}\chi'(i) \tag{17}$$

Therefore, the similarity formula for the category diagram G_i and the text graph G_j is

$$S2(G_i,\mathrm{G}_j)=\frac{\sum_{i\in V_{mcs}}\chi'(i)+\sum_{l\in L}\overline{\chi}'_l}{\max\{|G_i|,|G_j|\}} \tag{18}$$

EXPERIMENTS AND RESULTS ANALYSIS

In the Windows XP environment, using Eclipse 3.2 and JDK 1.5 as the development platform, the algorithm was written and realized. Experimental machine configuration is for P1.6GHz, 512MB memory, 120GB hard disk. Experimental data derived from Fudan University, Li Ronglu with the Chinese text classification corpus.

Experiment content

In order to verify text classification effectiveness based on the graph space model with the performance evaluation parameters: recall, precision and F1 value.

Experimental data from the data source selected three texts of sports, economy and art, the training set is 3 ×700 texts, test set is for the sports 1254, economy 1601, art 850. community discovery algorithm for the extraction of three categories, each of the 1000 characters are composed of three classes graphs, the tested text expressed as graph is divided into the greatest category similarity With the category graph. Experiments were done classification performance analysis on the following two similarity formula, $S1(G_1,G_2)$ is similarity formula determined by the MCS, it is as follows:

$$S1(G_1,G_2)=\frac{|mcs(G_1,G_2)|}{\max\{|G|,|G_2|\}} \quad \textbf{(19)}$$

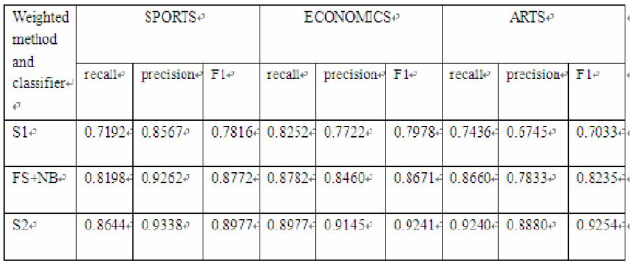
And based on the structural equivalence theory, the formula (22) was improved, it is changed into formula (23).

$$S1(G_1,G_2)=\frac{|mcs(G_1,G_2)|}{\max\{|G|,|G_2|\}} \quad \textbf{(20)}$$

The experimental results and analysis

Experiment counted two similarity formula for classification of recall, precision and F1 values in TABLE 1and in Figure 1, joined by the experiment[21] calculation the feature weights and classification effect in NB classification are compared.

TABLE 1 : The recall, precision and F1 value of three methods in each category



Weighted method and classifier	SPORTS			ECONOMICS			ARTS		
	recall	precision	F1	recall	precision	F1	recall	precision	F1
S1	0.7192	0.8567	0.7816	0.8252	0.7722	0.7978	0.7436	0.6745	0.7033
FS+NB	0.8198	0.9262	0.8772	0.8782	0.8460	0.8671	0.8660	0.7833	0.8235
S2	0.8644	0.9338	0.8977	0.8977	0.9145	0.9241	0.9240	0.8880	0.9254

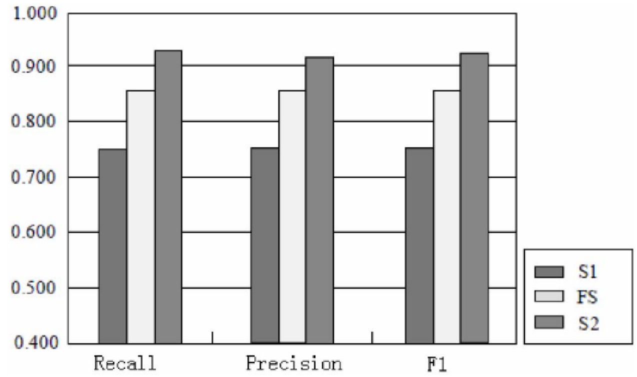


Figure 1 : Classification performance presentation

The recall, precision and F1 value for various categories, and macro average is made, the classification performance is as shown in Figure 1.

From the experimental statistical data, after joining common node and common chain structure equivalence analysis, between graphs similarity metric formulas S2 and S1 were compared, classification performance is improved to a certain extent.

CONCLUSION

Because of VSM lost text structure information, text classification method based on the space model is Launched study. Although using maximum common subgraph measure similarity of two graph is a relatively common method, but these methods have not made full use of space model containing lots of semantic information, text classification results were generally poor, therefore, on the structural equivalence basis, thus proposed the similarity measure formula based on graph space model, Finally results show the effectiveness of this method by experiment. The next step is how to give the similarity metrics for text classification based on the auxiliary dictionary text concept graph model. Which text classification approach to use depends on the requirements of subject.

THE FURTHER WORK AND THE PROBLEMS

With text information scale expansion and diversification of the forms text information field is full of opportunities and challenges, both in theory and in practice, for the study of text classification has great development space, different sample sets, model of text representation, feature selection algorithms, classification model and the methods of evaluation, which have a certain impact for the text classification results. The following research work mainly includes:

The difference of different words with the same text contribution to the theme of the text is very big, how to use prior knowledge to analyse semantic information of the different words contain, which will have the expression text for text semantic model is more accurate than the subject content.

Most researchers only to establish their own sets of documents and manual identification of each categories of documents, consumes a lot of time and effort at the same time, it is difficult to objectively evaluate their own work. So the research on Chinese text classification and research organization, is a urgent task of Chinese classified sample set.

ACKNOWLEDGEMENTS

The work was supported in part by a grant from Military and Civilian Integration Resarch Fund Project of Shaanxi Province (No:12JMR19), the Department of Scientific Research Fund Project of Shaanxi Province (No:2012JM8048) and the Scientific Research Fund Project of Weinan Normal University (No:13YKP015).

REFERENCES

[1] H.P.Luhn; Auto-encoding of documents for information retrieval systems[M]. Modern Trends in Documentation. London, England: Pergamon Press.1959:45-58

[2] M.E.Maron, J.L.Kumns; On Relevance, Probabilistic Indexing and Information Retrieval[J]. Journal of ACM., **7(3)**, 216-244 **(1959)**.

[3] G.Salton, M.J.Mcgill; An Introduction to Modem Information Retrieval[M]. Mcgraw-Hill, **(1983)**.

[4] Svetl ana Hensman; Construction of conceptual graph rep resentation of texts [C]. Proceeding s of the Student Research Work shop at HLT-NAACL, Bost on, 49-54 **(2004)**.

[5] Bhoopesh, Pushpak; Text clustering using semantics [C]. The11th International World Wide Web Conference, (WWW2002), Hawai, USA : 2002, **(2002)**.

[6] H.Uchida, M.Zhu; Senta TDella. UNL: A gift or a millennium[R]. Technical Report, The United Nations University, **(2000)**.

[7] Inderjeet Mani; Er ic bl oedorn : mult i-document sum mari zat ionby graph s earch and mat ching [C].Proceedings of the Fifteenth National Conferenceon Articial Intelligence, 622-628 **(1997)**.

[8] A.Schenker, M.Last, H.Bunk e et al.; Kandel: clustering of w ebdocument s u sing a graph m odel [A]. Web Document An alys is:Chall enges and Opportunit ies ed s[C]. Sigapore: World Scientific, 3-18 **(2003)**.

[9] Sanguthevar Rajasekaran,Vamsi Kundeti; Spectrum based techniques for graph isomorphism, International Journal of Foundations of Computer Science, **3(20)**, 479-499 **(2009)**.

[10] Jacobo Toran; Reductions to Graph Isomorphism, Theory of computing systems, **1(47)**, 288 **(2010)**.

[11] Ali Idarrou, Driss Mammass; An Approach based on Semantic Sub-graph Isomorphism, International Journal of Computer Applications, **1(51)**, 14-21 **(2012)**.

[12] Leander Schietgat, Fabrizio Costa, Jan Ramon, Luc De Raedt; Effective feature construction by maximum common subgraph sampling, Machine learning, **2(83)**, 137-16 **(2011)**.

[13] J.Mohr, B.Jain, A.Sutter, A.T.Laak, T.Steger-Hartmann, N.Heinrich, K.Obermayer; A maximum common subgraph kernel method for predicting the chromosome aberration test, Journal of chemical information and modeling, **10(50)**, 1821-1838 **(2010)**.

[14] Mirtha-Lina Fernandez, Gabriel Valiente; A graph distance metric combining maximum common subgraph and minimum common supergraph, Pattern recognition letters, **6(22)**, 753-758 **(2001)**.

[15] A.Kevin; Naude, Marking student programs using graph similarity, Computers & education, **2(54)**, 545-561 **(2010)**.

[16] G.Paul; Mezey, Graph representations of molecular similarity measures based on topological resolution, Journal of computational methods in sciences and engineering, **1(5)**, 109-114 **(2005)**.

[17] A.I.Gitelman, A.Herlihy; Isomorphic chain graphs for modeling spatial dependence in ecological data, Environmental and ecological statistics, **1(14)**, 27-40 **(2007)**.

[18] Mariam Daoud, Lynda Tamine, Mohand Boughanem; A personalized search using a semantic distance measure in a graph-based ranking model, Journal of Information Science, **12(37)**, 614 - 636 Dec **(2011)**.

[19] N.Davis, C.Giraud-Carrier, D.Jensen; A topological embedding of the lexiconfor semantic distance computation, Natural language engineering, **3(16)**, 245-275 **(2010)**.

[20] H.S.Min, J.Y.Choi, W.De Neve, Y.M.Ro; Near-Duplicate Video Clip Detection Using Model-Free Semantic Concept Detection and Adaptive Semantic Distance Measurement, IEEE Transactions on Circuits and Systems for Video Technology, **22(8)**, 1174-1187 **(2012)**.

[21] Xiaoqiang Jia; A Text Classification Algorithm based on the Community Discovery, International Review on computers and softwares, **7(3)**, 1303-1307 July **(2012)**.

[22] Xiaoqiang Jia, Jiangyan Sun; An Improved Text Classification method based on Gini Index, Journal of Theoretical and Applied Information Technology, **43(2)**, 267-273 July **(2012)**.