Journal of Theoretical and Applied Information Technology
28th February 2013. Vol. 48 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN TEXT CLASSIFICATION APPROACH BASED ON THE GRAPH SPACE MODEL

[1]XIAOQIANG JIA,

[1]School of Mathematics and Information Science, Weinan Normal University, Weinan 714000, Shaanxi, China

E-mail: 394892738@qq.com

ABSTRACT

To do the text classification on the basis of VSM, and use the maximum common subgraph to measure two graphs' similarities are the relatively common methods, but these methods have not made full use of lots of semantic information spatial model contained, so the text classification performance is generally poor. In order to improve the classification results of the graph, on the basis of the structural equivalence, this paper further analyzes the maximum common substructure graph nodes and edges if it is a true semantic equivalence, and puts forward a kind of improvement text similarity metrics based on the graph space model. Then apply it to the text classification, the classification performance has been improved. Finally, verify the effectiveness of this method by experiment.

Keywords: *Structural Equivalence， Text Classification, Graph Space Model, Maximum Common Subgraph, Similarity*

1. INTRODUCTION

The Internet is developing at a higher speed more than most people can imagine. Scholars pay close attention to how to find useful information and knowledge from Internet. For an ordinary Internet user, he tends to only use 1% amount of information on the Internet, which causes the waste of time and energy. In addition, compared with the web data, the traditional data mining research mainly focuses on data in database, and the data in the database is completely structured data, while the data on the web is semi structured, even isomerism. Obviously, the web data mining is more complex than database in data mining. In this context, Web data mining[1] is emerged as the times requiring, how fast, accurate and timely to meet user's need for information and knowledge become the target of web data mining. Since the concept of Web data mining is proposed by many scholars, they launched an in-depth discussion and research from their field of study with different theory and approaches. Text mining is an important research content of the current web data mining. The Internet is presented the hypertext form to the user, a webpage contains a variety of different data types, such as news reports, e-mail, pictures, books and multimedia, among this data, the number and size of text data is the maximum. As the information service quality requirements of users increasing, the traditional text management tools have been unable to be adapted to the massive text data processing requirements. Therefore, the text mining technology research is a problem worth discussing. Text classification is a key technology in the web text mining, including data mining, machine learning, neural networks, statistical and natural language processing and many other fields of study, and in information retrieval, information extraction, information filtering, automatic indexing, document organization and so on, it has a wide range of applications. In the past 40 years, domestic and foreign scholars on the text classification technology conducted in-depth research, and obtain many results attracting people's attention.

Foreign research in text classification can be roughly divided into three stages: the first stage (1958-1964) is mainly from dynamic classification of feasibility study, the second stage (1965-1974) is for automatic classification of experimental study, the third stage (1975- date) is the practical stage. Chinese text classification technology research began relatively late, generally began in the early nineteen eighties. The most frequently used are mostly based on the VSM. VSM in knowledge expression has a huge advantage, not only VSM is very simple, intuitive, and many of the statistical

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

information can be reflected through the VSM model, such as term frequency, inverse document frequency and the decision information. But in the construction of the VSM process, a lot of text structure information is lost, including the sequence between terms, terms appearing in the text of the position and the distance between the two terms and other information. In fact, the text structure information in natural language text is crucial, the different organization structure is likely to have a completely different semantic. In order to overcome the defects of VSM, many scholars put forward text representation methods based on graph model, such as Svetlana was proposed based on the auxiliary dictionary Verb Net and Word Net text concept graph representation, but these methods are too complex, it is difficult to give a similarity metric standard, therefore, the relatively simple graph space model[6,7] is discussed and researched.

VSM[3] is not only simple in form, and easy to operate. It is a representation model with more common useless and good effectiveness. In the construction of the VSM process, VSM losts a lot of text structure information, including the sequence between terms appearing, the position of terms appearing in the text and the distance information between terms. Text categorization[2] based on graph model are more complicated, it is difficult to give a similarity metric standard, although the maximum common subgraph measure similarity between two graph are relatively common methods, but these methods have not made full use of graph space model [7,8]containing large amounts of semantic information, so the text classification performance is generally poor.

As 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. 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. It is applied to text categorization, and the experimental results show text classification method based on the graph space model is feasible and effective.

Section 2 presents definitions and symbols related to graph models, In section 3, the distance measurement method between graphs will be introduced. Section 4 presents maximum common subgraph computations. On the basis of the above, the main problem discussed in the text is put forward in Section 5. Section 6 gives the definition of the common node structural equivalence and the common chain structure equivalence between the graphs. Section 7 proposes the improved data structure of graph and the new the similarity formula, and uses it to text classification. Section 7 means the experimental results and analysis. Section 9 gives a conlusion to the whole paper.

2. DEFINITIONS AND SYMBOLS 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. In simple case, a node or edge is only one label. In the graph space model of text, the general node label information represents the node term, and edge label information reflects if the two node are 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 termed 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 are no special, node labeled terms, edges marked if the two node are 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

Journal of Theoretical and Applied Information Technology
28th February 2013. Vol. 48 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

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 exists 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.

Definition 5: G, G_1 and G_2 are graphs, 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.

Definition 6: G, G_1 and G_2 are graphs, if the subgraph of G_1 is isomorphism with G, and the subgraph of G_2 is isomorphism with G, then G is the common supergraph of G_1 and G_2.

Definition 7: G, G_1 and G_2 are graphs, if G is the minimum common supergraph[14], then it is marked with $MCS'(G_1,G_2)$, if G is the common supergraph of G_1 and G_2, there not exists other common supergraph G' of G_1 and G_2, then $|G'|<|G|$ holds.

3. 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 and the minimum common subgraph.

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

(2)MMCSN

$$d_{MMCSN}(G,G') = 1 - \frac{|mcs(G,G')|}{|MCS(G,G')|} \quad (2)$$

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

4. MAXIMUM COMMON SUBGRAPH COMPUTATIONS

By the formula (1)-(5) can be seen, the maximum common subgraph is the key problem calculated the distance between graph and graph, which relates to the graph isomorphism identification. At present, according to the domestic and foreign research progress of graph isomorphism, no exact graph isomorphism, inexact subgraph isomorphism and precise subgraph isomorphism has been shown to belong to NPC problem, and precise graph isomorphism is not classified P problem, and also not be classified NPC problem, which is one of the unsolved problems. Thus the further study is very necessary. Specific targeting of the definition of 1 expressed data structure in the text graph, because the node and terms are one-to-one correspondence, all edges are uniformly labeled when two nodes are adjacent, so the calculation of the maximum common subgraph complexity is $O(n^2)$. Therefore, how many labels whether text nodes in the graph or edge have, it is simplified the form of the data structure as the definition of 1 to solve the two largest common subgraphs, that is node label select node indicated by the term, edge label select information whether the two nodes are adjacency.

According to the definition of 1, set up G_1 and G_2 are pairs of graphs represented text, then the maximum common subgraph G_{mcs} of G_1 and G_2 is generated as follows:

(1) Search all public node set V_{mcs} expressed in the same term of G_1 and G_2, nodes of V_{mcs} is as nodes of G_{mcs}.

(2) Check all node pairs in V_{mcs} by the step of (1), if the node pairs are on the presence of adjacent relation, then add the edge to public set of edges E_{mcs}, the edges in V_{mcs} is as the edges in G_{mcs}. The improved data structure of graph for the definition of 11, experiments for the maximum common subgraph will also use the same method.

5. INTRODUCING THE PROBLEM

Use the maximum common subgraph to measure a graph similarity[15-16], it is very intuitive and consistent with the conventional thinking. In addition to graph edit distance method, the (1)-(5) several formulas are dependent on the maximum common subgraph to calculate the distance between graph and graph. However, in the formula "$|mcs(G_1,G_2)|$" is simple statistical the number of the common nodes and common edges of the maximum

Journal of Theoretical and Applied Information Technology
28th February 2013. Vol. 48 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

common subgraph, but no deep take into account representative terms of these common nodes and the common edges and the adjacent relation between terms and terms, whether or not in different segments indicate the same semantic.

6. STRUCTURAL EQUIVALENCE

6.1 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:

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

An isolation term generally does not reflect certain semantic information, and 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 the undirected graph, so the row and column of matrix reflect 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 computes 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 needed 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:

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

B' is defined as follows:

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

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 $\xi'=(A'_{i,0},A'_{i,1},\cdots,A'_{i,k-1})$ and $\eta'=(B'_{i,0},B'_{i,1},\cdots B'_{i,k-1})$, which is composed of A' and B'. Firstly, the mean value and variance for row vector of the adjacency matrix are as follows:

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

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

$$\mu_{\eta'}=\frac{1}{n}\sum_j B'_{i,j} \quad (8)$$

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

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

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

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 generates in the process of graph expansion, but this kind of nodes in G_1 do not exist, so there is no necessary to do the node similarity.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Moreover, since it has certain particularity in the text graph model, 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:

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

6.2 Common Chain Structure Equivalence

The Formula (14) only considers the node structure equivalent degree, in text the phrase structure also contains 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 to constitute the minimum connected subgraph of graph that 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 to 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 that is the 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:

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

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

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

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:

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

7. THE IMPROVED GRAPH 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.

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

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

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 be $(w_1, w_2, \cdots, w_n)$, each weight proportion

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

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:

$$\alpha_2(j)=x_1w_1(j)+x_2w_2(j)+\cdots+x_nw_n(j) \quad (15)$$

If the training set is web formatted text, in text terms location information is 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 appears in the other positions, the node weight is 2. If a term appears 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, the training text set constructs 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 are 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:

$$\chi'(i)=\chi(i)\alpha_2(i) \quad (16)$$

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

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

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

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

8. EXPERIMENTS AND RESULTS ANALYSIS

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

8.1 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:

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

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

$$S2(G_1,G_2)=\frac{\sum_{i\in V_{mcs}}\chi'(i)+\sum_{l\in L}\overline{\chi}_l'}{\max\{|G_1|,|G_2|\}} \quad (20)$$

8.2 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.

Journal of Theoretical and Applied Information Technology
28th February 2013. Vol. 48 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

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

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.

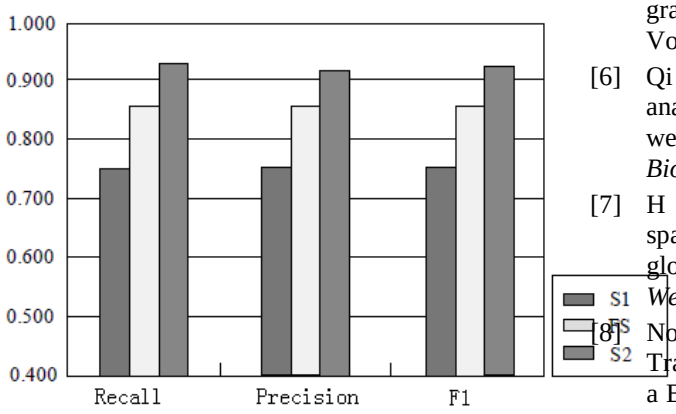

Figure 1. Classification Performance Presentation

9. CONCLUSION

Because of VSM lost text structure information, text classification method based on the space model is studied. Although to use maximum common subgraph to measure similarity of two graph is a relatively common method, but it has not made full use of space model containing lots of semantic information, and the text classification results were generally poor. Therefore, on the structural equivalence basis, this paper 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. To choose which text classification approach depends on the requirements of subject.

REFERENCES

[1] Olaru C., Wehenkel L., Data mining, *IEEE Computer Applications in Power*, Vol.3,No.12,1999, pp.19-25.

[2] Jiangning Wu , Zhaoguo Xuan ,Donghua Pan, Enhancing Text Representation For Classification Tasks with Semantic Graph Structures, *International journal of innovative computing, information and control*, Vol.5 B,No.7,2011,pp.2689-2698.

[3] Toru IshibashiS, Yasufumi Takama, Proposal of M2VSM and Its Comparison with Conventional VSM, *Artificial Intelligence and Knowledge Based Processing*, Vol.485,No.104,2004, pp.1-6.

[4] Jukna S, On graph complexity, Combinatorics, *probability & computing*, Vol.6, No.15,2006,pp.855-876.

[5] Akio Kawauchi, On a complexity of a spatial graph, *Study on physical properties*, Vol.1,No.92,2009, pp.111-114.

[6] Qi Zh , Li L ,Zhang Zm ,Qi Xq, Self-similarity analysis of eubacteria genome based on weighted graph, *Journal of Theoretical Biology*, Vol.1,No.280,2011,pp.10-18.

[7] H Deng ,M R Lyu ,I King, Effective latent space graph-based re-ranking model with global consistency, *Proc 2nd ACM Int Conf Web Search Data Mining*,2009,PP.212-221.

[8] Noor ainy harish,Razidah ismal, Tahir ahmad, Transformation of Fuzzy State Space Model of a Boiler System: A Graph Theoretic Approach, *WSEAS Transactions on Mathematics*, Vol.7,No.9,2010,pp.669-678.

[9] Sanguthevar Rajasekaran,Vamsi Kundeti, spectrum based techniques for graph isomorphism, *International Journal of Foundations of Computer Science*, Vol.3, No.20,2009,pp.479-499.

[10] Jacobo Toran, Reductions to Graph Isomorphism, *Theory of computing systems*, Vol.1,No.47,2010, pp.288.

[11] Ali Idarrou, Driss Mammass, An Approach based on Semantic Sub-graph Isomorphism, *International Journal of Computer Applications* , Vol.1,No.51, 2012,pp.14-21.

[12] Leander Schietgat ,Fabrizio Costa, Jan Ramon , Luc De Raedt, Effective feature construction by maximum common subgraph sampling, *Machine learning*, Vol.2,No.83, 2011,pp.137-16.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[13] Mohr, J. , Jain, B. ,Sutter, A., Laak, A.T. ,Steger-Hartmann, T.,Heinrich, N. ,Obermayer, K., A maximum common subgraph kernel method for predicting the chromosome aberration test, *Journal of chemical information and modeling*, Vol.10, No.50, 2010,pp.1821-1838.

[14] Mirtha-Lina Fernandez , Gabriel Valiente, A graph distance metric combining maximum common subgraph and minimum common supergraph, *Pattern recognition letters*,Vol.6, No.22,2001,pp. 753-758.

[15] Kevin A. Naude, Marking student programs using graph similarity, *Computers & education*, Vol.2, No.54, 2010,pp.545-561.

[16] Paul G. Mezey, Graph representations of molecular similarity measures based on topological resolution, *Journal of computational methods in sciences and engineering*, Vol.1, No.5, 2005,pp. 109-114.

[17] Gitelman AI,Herlihy A, Isomorphic chain graphs for modeling spatial dependence in ecological data, *Environmental and ecological statistics*, Vol.1, No.14, 2007,pp.27-40.

[18] Mariam Daoud, Lynda Tamine, and Mohand Boughanem, A personalized search using a semantic distance measure in a graph-based ranking model, *Journal of Information Science*, Vol.12, No. 37, Dec 2011,pp. 614 - 636.

[19] N. DAVIS,C. GIRAUD-CARRIER ,D. JENSEN, A topological embedding of the lexiconfor semantic distance computation, *Natural language engineering*, Vol. 3, No. 16, 2010, pp. 245-275.

[20] Min, H.-S. ; Choi, J. Y. ; De Neve, W. ; Ro, Y. M, 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*, Vol.22, No. 8,2012,pp.1174-1187.

[21] Xiaoqiang Jia, A Text Classification Algorithm based on the Community Discovery,*International Review on computers and softwares*,Vol. 7, No. 3,July 2012, pp.1303-1307.