|  |  |  |
| --- | --- | --- |
|  | Available online at www.sciencedirect.com |  |
|  |
| AASRI Procedia 3 ( 2012 ) 680 – 685 |

2012 AASRI Conference on Modeling, Identification and Control

Research On the Selection of Feature Transfer Relations in Latent Semantic Indexing

Dongyang Jianga,\* , Wei Zhengb

*aInformation Engineering Department, Liaoning Jidian Polytechnic,Dandong,118009, China b Alibaba Group,Hangzhou, 310000,China*

**Abstract**

The latent semantic index (LSI) has been widely used in many fields of natural language processing in which co-occurrence features can be captured by the transfer relations between the documents and in the documents. Document features with a higher frequency in the collection of the document are more likely to introduce some unreasonable feature transfer relations to the latent semantic space which affects the similarity between features and between documents in document sets in our recent study. In the paper a feature optimize technology in latent semantic indexing that uses feature transfer relation in documents and between documents is proposed. By the complete-link algorithm, the experimental results show that the method effectively improves the performance of latent semantic indexing.

*Keywords:* Latent semantic index, feature transfer, feature similar matrix;

**1. Introduction**

With the development of information technology, a lot of document resources are needed that helps the discovery of the theme, information retrieval, and so on. Therefore, text clustering technology came into being. It is a very important part of natural language processing. Text clustering technique made great success in document clustering. There are a large number of synonyms, near-synonym and other unique natural language phenomena in document clustering. We will use LSI to explore and resolve these linguistic phenomena to improve the performance of document clustering in the paper.

**2.The feature transfer relations**

\*e-mail: linda164@163.com

2212-6716 © 2012 The Authors. Published by Elsevier B.V. Open access under [CC BY-NC-ND license.](http://creativecommons.org/licenses/by-nc-nd/3.0/) Selection and/or peer review under responsibility of American Applied Science Research Institute doi: 10.1016/j.aasri.2012.11.108

*Dongyang Jiang and Wei Zheng / AASRI Procedia 3 ( 2012 ) 680 – 685*  681

The co-occurrence information of the terms can be captured when Singular Value Decomposition (SVD) is decomposed proposed by [1]. The example preferred in [2] is as shown by table 1 and table 2 is the feature document matrix. The term weight represents word frequency of the term and the matrix is decompose by singular value dropping to a two-dimensional space. By the comparison between the similarity matrix decomposed shown by Table 3 and the one undecomposed, the similarity weights of the table 3 made obvious changes. In table 2 the similarity between the terms is zero, but there does not exist a value of 0 which means co-occurrence information of some terms is improved and some one is weakened. The similarity of 0 is between some terms undecomposed that means there is no or little relations. The undecomposed similarity of the user (t4) and human (t1) is 0 but decomposed similarity is 1.0003. By the changes of the similarity value, we can think "user" and "interface", "interface" and "human", "user" and "human" co-occur. In LSI "user" and "human" projected to the same dimension space.

Table 1. The technology Memorandum titles

|  |  |
| --- | --- |
| number | article title |
| c1  c2  c3  c4  c5  m1  m2  m3  m4 | Human machine interface for Lab ABC computer applications A survey of user opinion of computer system response time The EPS user interface management system  System and human system engineering testing of EPS  Relation of user-perceived response time to error measurement The generation of random, binary, unordered trees  The intersection graph of paths in trees  Graph minors IV: Widths of trees and well-quasi-ordering Graph minors: A survey |

Table 2. Deerwester feature document matrix

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | c1 c2 | | c3 | c4 | c5 | m1 | m2 | m3 | m4 |
| human(t1) | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| interface(t2) | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| computer(t3) | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| user(t4) | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| system(t5) | 0 | 1 | 1 | 2 | 0 | 0 | 0 | 0 | 0 |
| response(t6) | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| time(t7) | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| EPS(t8) | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| survey(t9) | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| trees(t10) | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 |
| graph(t11) | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| minors(t12) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| X(t13) | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

The degree of similarity between features reflects the correlation of between the terms. The weight value not only reflects the correlation between the features but also embodies the co-occurrence information between the features in SVD space. As can be seen from Table 3, the similarity value of "time (t7)" and "graph(t11)" is 0.4988. These terms are from different classes and the variation of the terms can be considered as the co-occurrence of "time" and "user". Assuming in these nine articles, one common feature is in each document. In this semantic space generated by the document collection, for the mutual transmission between the terms, some feature co-occurrence information that not exists appear in the document so that some non-existent feature co-occurrence information which is noise data will generate between the documents. Fox example,a term X is added to the each document of Table 1 whose feature weight value is 1. Feature

682  *Dongyang Jiang and Wei Zheng / AASRI Procedia 3 ( 2012 ) 680 – 685*

document matrix in Table 1 are decompose by SVD and the similarity values between the features are gotten by the Equation (1) and whose similarity matrix is shown by Table 4. From the Table 4, the weight value of "compute(t3)", "response(t6)" and "time(t7)" are all 0.6925, compared with the corresponding ones in Table 4 that the weight value is weakened. As seen from Table 1, the common feature X added makes the weight value of these words weakened that should be very near.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *A* | *KT* | *A* | *K* | ( | *T* | *K* | *S* | *K* | *D* | *K* | *T* | ) | *T* | *T* | *K* | *S* | *K* | *D* | *K* | *T* | *D* | *K* | *S* | *K* | *T* | *T* | *K* | *T* | *T* | *K* | *S* | *K* | *D T K* | *K* | *T* | *D* | *K* | *S* | *K* | ( | *D* | *K* | *S* | *K* | ) | *T* | *(1)* |

Table 3. Deerwester feature similarity matrix truncated to two-dimension

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| t1 | | t2 | t3 | t4 | t5 | t6 | t7 | t8 | t9 | t10 | t11 | t12 |
| t1 | 0.5987 | 0.5113 | 0.5980 | 1.0003 1.6996 | | 0.6102 | 0.6102 | 0.7980 | 0.2994 | -0.2987 | -0.4004 | -0.3003 |
| t2 | 0.5113 | 0.5002 | 0.4987 | 0.9001 | 1.5002 | 0.4998 | 0.4998 | 0.6909 | 0.2988 | -1.9999 | -0.1984 | -0.0984 |
| t3 | 0.5980 | 0.4986 | 0.7101 | 0.9978 | 2.1001 | 0.6993 | 0.6993 | 0.8011 | 0.5984 | 0.2010 | 0.3002 | 0.1987 |
| t4 | 1.0003 | 0.9001 | 0.9978 | 2.0011 | 3.0140 | 1.1908 | 1.1908 | 1.2908 | 0.9971 | 0.1995 | 0.4002 | 0.2998 |
| t5 | 1.6996 | 1.5002 | 2.1001 | 3.0140 | 5.0024 | 2.0121 | 2.0121 | 1.9987 | 0.9932 | -0.3986 | -0.3765 | -0.3004 |
| t6 | 0.6102 | 0.4998 | 0.6993 | 1.1908 | 2.0121 | 0.9011 | 0.9011 | 0.8012 | 0.8132 | 0.4014 | 0.4988 | 0.3989 |
| t7 | 0.6102 | 0.4998 | 0.6993 | 1.1908 | 2.0121 | 0.9011 | 0.9011 | 0.8012 | 0.8132 | 0.4014 | 0.4988 | 0.3989 |
| t8 | 0.7980 | 0.6909 | 0.8011 | 1.2908 | 1.9987 | 0.8012 | 0.8012 | 1.1411 | 0.3982 | -0.3991 | -0.3999 | -0.2989 |
| t9 | 0.2994 | 0.2988 | 0.5984 | 0.9971 | 0.9932 | 0.8132 | 0.8132 | 0.3982 | 1.0001 | 0.9001 | 1.2006 | 0.9020 |
| t10 -0.2987 | | -1.9999 | 0.2010 | 0.1995 | -0.3986 | 0.4014 | 0.4014 | -0.3991 | 0.9001 | 1.6017 | 2.0200 | 1.3976 |
| t11 -0.4004 | | -0.1984 | 0.3002 | 0.4002 | -0.3765 | 0.4988 | 0.4988 | -0.3999 | 1.2006 | 2.0200 | 3.0002 | 2.0013 |
| t12 -0.3003 | | -0.0984 | 0.1987 | 0.2998 | -0.3004 | 0.3989 | 0.3989 | -0.2989 | 0.9020 | 1.3976 | 2.0013 | 1.2995 |

Table 4. Feature similarity matrix truncated to two-dimension with the features

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| t1 | | t2 | t3 | t4 | t5 | t6 | t7 | t8 | t9 | t10 | t11 | t12 | t13 |
| t1 | 0.6698 | 0.6058 | 0.6540 | 1.0823 | 1.7542 | 0.6878 | 0.6878 | 0.8530 | 0.4619 | -0.0092 | -0.0270 | -0.0190 1.7621 | |
| t2 | 0.6058 | 0.5529 | 0.6017 | 0.9868 | 1.5728 | 0.6323 | 0.6323 | 0.7620 | 0.4562 | 0.0923 | 0.0895 | 0.0615 | 1.7377 |
| t3 | 0.6540 | 0.6017 | 0.6595 | 1.0731 | 1.6847 | 0.6925 | 0.6925 | 0.8135 | 0.5296 | 0.1966 | 0.2063 | 0.1426 | 2.0148 |
| t4 | 1.0823 | 0.9868 | 1.0731 | 1.7615 | 2.8124 | 1.1277 | 1.1277 | 1.3631 | 0.8078 | 0.1460 | 0.1385 | 0.0951 | 3.0775 |
| t5 | 1.7542 | 1.5728 | 1.6847 | 2.8124 | 4.6312 | 1.7733 | 1.7733 | 2.2596 | 1.1051 | -0.2976 | -0.3803 | -0.2658 4.2230 | |
| t6 | 0.6878 | 0.6323 | 0.6925 | 1.1277 | 1.7733 | 0.7272 | 0.7272 | 0.8565 | 0.5527 | 0.1958 | 0.2046 | 0.1413 | 2.1032 |
| t7 | 0.6878 | 0.6323 | 0.6925 | 1.1277 | 1.7733 | 0.7272 | 0.7272 | 0.8565 | 0.5527 | 0.1958 | 0.2046 | 0.1413 | 2.1032 |
| t8 | 0.8530 | 0.7620 | 0.8135 | 1.3631 | 2.2596 | 0.8565 | 0.8565 | 1.1041 | 0.5158 | -0.200 | -0.2486 | -0.1734 1.9730 | |
| t9 | 0.461 | 0.4562 | 0.5296 | 0.8078 | 1.1051 | 0.5527 | 0.5527 | 0.5158 | 0.6124 | 0.7624 | 0.8516 | 0.5907 | 2.3160 |
| t10 -0.009 | | 0.0923 | 0.1966 | 0.1460 | -0.297 | 0.1958 | 0.1958 | -0.200 | 0.7624 | 2.0108 | 2.2765 | 1.5803 | 2.8554 |
| t11 -0.027 | | 0.0895 | 0.2063 | 0.1385 | -0.380 | 0.2046 | 0.2046 | -0.2486 | 0.8516 | 2.2765 | 2.5777 | 1.7894 | 3.1887 |
| t12 -0.019 | | 0.0615 | 0.1426 | 0.095 | -0.265 | 0.1413 | 0.1413 | -0.1734 | 0.5907 | 1.5803 | 1.7894 | 1.2422 | 2.2119 |
| t13 1.7621 | | 1.7377 | 2.0148 | 3.0775 | 4.2230 | 2.1032 | 2.1032 | 1.9730 | 2.3160 | 2.8554 | 3.1887 | 2.2119 | 8.7596 |

The similarity degree between the documents mainly depends on the number of the co-occurrence features. In the generating latent semantic space, because of the transitivity between the features, the latent relations will be excavated. There is perhaps high similarity between the documents whose similarity are little or non. The similarity between the documents is calculated by Equation (1) and Table 6 is the similarity matrix after being adding the feature X. As seen from the data of the matrix, a clear distinction exists between the documents. The same type of documents has a higher similarity and different type of documents has a lower similarity. For example, there is a high similarity between "M4" and "c2" whose value is 1.1213. As these documents all have the term "survey", the "survey" weight value of co-occurrence is strengthened.

*Dongyang Jiang and Wei Zheng / AASRI Procedia 3 ( 2012 ) 680 – 685*  683

Table 5. similarity matrix between the documents

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | c1 | c2 | c3 | c4 | c5 | m1 | m2 | m3 | m4 |
| c1 | 0.4551 | 1.2753 | 1.0659 | 1.2781 | 0.5772 | -0.0613 | -0.1259 | -0.1690 | -0.0109 |
| c2 | 1.2753 | 4.2759 | 2.9949 | 3.4188 | 2.0045 | 0.2321 | 0.5674 | 0.8213 | 1.1213 |
| c3 | 1.0659 | 2.9949 | 2.4965 | 2.9916 | 1.3562 | -0.1389 | -0.2845 | -0.3812 | -0.0124 |
| c4 | 1.2781 | 3.4188 | 2.9916 | 3.6272 | 1.5312 | -0.2656 | -0.5671 | -0.7749 | -0.2975 |
| c5 | 0.5772 | 2.0045 | 1.3562 | 1.5312 | 0.9454 | 0.1449 | 0.3477 | 0.4996 | 0.6212 |
| m1 | -0.0613 | 0.2321 | -0.1389 | -0.2656 | 0.1449 | 0.2404 | 0.5461 | 0.7674 | 0.6637 |
| m2 | -0.1259 | 0.5674 | -0.2845 | -0.5671 | 0.3477 | 0.5461 | 1.2410 | 1.7440 | 1.5125 |
| m3 | -0.1690 | 0.8213 | -0.3812 | -0.7749 | 0.4996 | 0.7674 | 1.7440 | 2.4509 | 2.1280 |
| m4 | -0.0109 | 1.1213 | -0.0124 | -0.2975 | 0.6212 | 0.6637 | 1.5125 | 2.1280 | 1.8892 |

Table 6. Similarity matrix after being adding feature X

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | c1 | c2 | c3 | c4 | c5 | m1 | m2 | m3 | m4 |
| c1 | 1.3942 | 2.5358 | 2.1039 | 2.2975 | 1.5799 | 0.9054 | 1.0610 | 1.1687 | 1.2278 |
| c2 | 2.5358 | 4.8108 | 4.0839 | 4.5541 | 2.8929 | 1.3060 | 1.3568 | 1.3916 | 1.6655 |
| c3 | 2.1039 | 4.0839 | 3.5082 | 3.9534 | 2.4091 | 0.9245 | 0.8585 | 0.8122 | 1.1170 |
| c4 | 2.2975 | 4.5541 | 3.9534 | 4.4958 | 2.6400 | 0.8475 | 0.6649 | 0.5378 | 0.9497 |
| c5 | 1.5799 | 2.8929 | 2.4091 | 2.6400 | 1.7923 | 0.9930 | 1.1468 | 1.2532 | 1.3363 |
| m1 | 0.9054 | 1.3060 | 0.9245 | 0.8475 | 0.9930 | 1.1735 | 1.6730 | 2.0198 | 1.7724 |
| m2 | 1.0610 | 1.3568 | 0.8585 | 0.6649 | 1.1468 | 1.6730 | 2.4612 | 3.0084 | 2.5732 |
| m3 | 1.1687 | 1.3916 | 0.8122 | 0.5378 | 1.2532 | 2.0198 | 3.0084 | 3.6948 | 3.1290 |
| m4 | 1.2278 | 1.6655 | 1.1170 | 0.9497 | 1.3363 | 1.7724 | 2.5732 | 3.1290 | 2.7052 |

**3. Experiments**

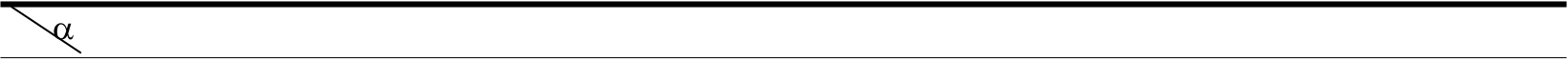
*3.1 corpus*

The corpus Tancorpv 1.0 used in the experiment is from the Chinese Academy of Sciences , Dr Tan Songbo and the text classification corpus from Sogou Lab. 12 classes are randomly selected from the 12 categories in the Tancorpv 1.0 which is 2,400 texts in all named as the Chinese Academy of Science Corpus 1 , the smallest text is 1kb, and the largest one is 14.7kb. One thousand texts are randomly selected in 9 classes from the text corpus of the Sogou Lab whose largest class contains 200 documents and whose smallest class contains 80 documents. 3000 texts are randomly selected from 60 smaller classes named as the Chinese Academy of Science corpus 2.

*3.2 Evaluation*

The evaluation of the clustering effect in the paper refers to the evaluation methods in information retrieval and each clustering result is looked on as a result of the query so that for some ultimate clustering category r and the original scheduled class i, whose F-measure[3-4] precision and recall are defined as below:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *recall* | ( | *i r* , ) | *n i r* ( , ) / | | *in* | | *(2)*  *(3)*  *rn* is the text number of class r. ni is the |
| *precision* | | | (, ) | *n i r* ( , ) / | | *nr* |
| Wherein n(i,r) is text data in the clustering including class i. | | | | | | |



684  *Dongyang Jiang and Wei Zheng / AASRI Procedia 3 ( 2012 ) 680 – 685*

text number included by the class i. Thus, the F-measure of clustering r and clustering i is calculated as below:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Fi r* | ) | 2 |  | *recall* | ( , *i r* | ) |  | | *precision* | | ( , *i r* | | ) | *(4)* |
|  |  |  | *precision* | | ( , *i r* | | | ) |  | *recall* | ( , *i r* | ) |  |  |

F-measure of each class is the largest value of that category in all classes. According to the F-measure, the clustering performance is evaluated as below:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *MacroF* | 1 | *i* | *n* | *i* | max{ | *F i r* | )} | *(5)* |
| *n* | |

*3.3 The experiment and analysis*

The features are firstly selected on the corpus. For the different experimental corpus, different thresholds

×FT are set (FT is the total number of documents for each experimental corpus; is scale factor whose value in the range [0, 1] ; ×FT is rounded).  The feature of DFij >FT is filtered off, forming a new feature space. The feature weight is calculated by TF-IDF[5], by the filtering feature document matrix generated through vector space model, then decomposing the matrix by SVD [6]. In LSI space, the text similarity is computed by the calculation method of the vector angle cosine. The clustering is by Complete-link algorithm.

Table 7. Clustering performance of feature transfer relation selection on LSI Sogou corpus

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| K | 5 | 7.5 | 10 | 12.5 | 15 | 17.5 | 30 | 50 | 70 |
| 5 | 45.1217 | 43.6923 | 52.9435 | 48.4663 | 47.6739 | 52.6592 | 51.4016 | 50.0703 | 50.0703 |
| 10 | 57.033 | 59.6071 | 58.67 | 58.9622 | **65.3312** | 60.2906 | 62.5312 | 58.7542 | 58.7542 |
| 30 | 53.1776 | 49.9327 | 49.9102 | 55.3107 | 55.47 | 53.9627 | 47.2007 | 54.5837 | 54.5837 |
| 50 | 48.5399 | 48.5399 | 41.7378 | 45.3191 | 46.7387 | 45.5403 | 37.0255 | 47.7429 | 47.7429 |
| 100 | 33.4347 | 45.4616 | 32.5768 | 34.7154 | 32.7541 | 32.1541 | 36.0562 | 36.8885 | 36.8885 |
| 150 | 37.4087 | 35.0805 | 35.1398 | 44.0819 | 38.5151 | 37.5466 | 39.2708 | 37.3287 | 37.3287 |
| 200 | 34.0331 | 31.2605 | 33.8221 | 32.8115 | 35.4574 | 35.885 | 32.577 | 29.183 | 29.183 |

Table 8. Clustering performance of feature transfer relation selection on LSI Chinese Academy of Science corpus 1

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| K | 5 | 6 | 8.5 | 11 | 13.5 | 16 | 18.5 | 30 | 50 | 70 |
| 57.73 | 55.3971 | 58.6267 | 57.6593 | 60.8423 | 60.9406 | 56.5363 | 54.5817 | 54.5817 |
| 10 | 67.6855 | 70.8075 | 71.5772 | **75.8125** | 69.537 | 74.8436 | 71.2974 | 73.0197 | 73.0197 |
| 30 | 53.5434 | 53.9958 | 53.1723 | 51.6969 | 53.3206 | 56.312 | 49.8639 | 52.5935 | 52.5935 |
| 50 | 43.1301 | 39.053 | 45.259 | 43.1051 | 44.2841 | 41.4488 | 40.244 | 46.3784 | 46.3784 |
| 100 | 38.8545 | 42.198 | 38.7509 | 36.0271 | 43.8528 | 42.6761 | 42.9188 | 48.5965 | 48.5965 |
| 150 | 32.9213 | 38.852 | 32.6512 | 38.2174 | 33.7758 | 39.224 | 45.3645 | 34.7402 | 34.7402 |
| 200 | 29.0897 | 33.1234 | 35.5488 | 34.138 | 35.7931 | 35.0818 | 34.306 | 38.9424 | 38.9424 |

As seen from the experimental results of Table 8, clustering performance firstly ascends and then descends with the increasing of on Sogou corpus and the Chinese Academy of science corpus 1. On Sogou corpus, when is 0.40, the clustering performance is highest. When is 1, the clustering performance has similar states on the Chinese Academy of Science corpus 1. However, for the Chinese Academy of Science corpus 2,

*Dongyang Jiang and Wei Zheng / AASRI Procedia 3 ( 2012 ) 680 – 685*  685

the clustering performance ascends, lastly to be the highest. This shows that the selecting of the appropriate threshold and the filtering out features of the document frequency over the threshold can not only reduce the

|  |  |  |
| --- | --- | --- |
| dimension of the feature space but also improve the performance of the clustering. |  | 100% ) means the |

feature transfer relationship have not been selected. The F-measure value of the clustering results will not change any longer when the feature of document frequency less than 50% FT is as a new feature collection. From these three corpus, when the feature document frequency is reserved between 10% and 15%, some feature transfer relationship can be effectively filtered out in LSI space and unreasonable co-occurrence features and some noise data can be eliminated.

Table 9. Clustering performance of feature transfer relation selection on LSI Chinese Academy of Science corpus 2

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 8 10 12 13 14 15 30 50 70  5 28.1901 29.309 28.0013 29.0241 29.1364 28.0656 29.1304 26.3343 26.3343 | | | | | | | | | |
| 10 | 39.5704 | 39.2226 | 41.157 | 41.292 | 39.7076 | 39.6279 | 39.4548 | 37.9426 | 37.9426 |
| 30 | 54.6392 | 54.5375 | 54.4965 | 54.4377 | 55.1985 | 54.5102 | 56.2248 | 55.6878 | 55.6878 |
| 50 | 55.31 | 55.5375 | 54.9462 | 55.3946 | 56.6932 | 56.0832 | 55.1885 | 55.9218 | 55.9218 |
| 100 | 54.6585 | 56.3235 | 56.6641 | 56.1209 | 53.1921 | 54.5315 | 52.1837 | 54.8526 | 54.8526 |
| 110 | 55.9208 | 55.8911 | **58.9859** | 56.6732 | 55.9022 | 56.9077 | 56.8547 | 57.3901 | 57.3901 |
| 150 | 52.9771 | 53.5297 | 54.9087 | 53.5675 | 51.6918 | 54.2409 | 49.0994 | 50.8681 | 50.8681 |
| 200 | 47.1427 | 46.696 | 51.3688 | 50.0678 | 51.4326 | 51.9511 | 50.5097 | 48.5113 | 48.5113 |

**4. Conclusion**

In this paper, we think that the transfer number between features has a great impact on the performance of latent semantic indexing. As the feature transfer number increases, some non-existent feature co-occurrence information appear which affects the similarity between features so that affects the performance of the latent sematic indexing. Before the decomposing of SVD, the feature of document collection is selected by DF feature in order to reduce the feature transfer number and non-existent feature co-occurrence information. The DF method used by our paper can selected features with documents in document collection and simply filters the transfer number between features. The next step, we will study on the feature selection based on conditional entropy between the features and conditional entropy.

**References**

[1] A. Kontostathis, W. M. Pottenger. A mathematical view of latent semantic indexing: Tracing term co-occurrences [R]. Technical report, LU-CSE-02-006, Dept.of Computer Science and Engineering, Lehigh University, 2002   
[2] Deerwester S. C., Dumais S. T, Landauer T. K, et al. Indexing by latent semantic analysis [J]. Journal of the American Society of Information Science, 1990, 41(6):391-407   
[3] Lewin D D, Ringuuette M. A comparison of two learning algorithms for text categorization[C]//proc of the Third Annual Symposium on Document Analysis and Information Retrieval,1994:81-93   
[4] Joachines T.Text categorization with support vector machines:learning with many relevant features[C]//proc of the 10th Eurospeech Conference on Machine Learning(ECML),1998:137-142 [5] Wiemer-Hastings, P. How latent is latent semantic analysis? [A]. In: Proceedings of the Sixteenth International Joint Conference on Articial Intelligence [C], 1999:932-937   
[6] Ding C.H.Q. A similarity-based probability model for latent semantic indexing [A]. In Proceedings of the Twenty-second Annual International ACM/SIGIR Conference on Research and Development in Information