

文章编号: 1003-0077(2009)01-0069-08

# 潜在语义索引中特征优化技术的研究

季 铎,郑 伟,蔡东风

(沈阳航空工业学院 知识工程中心,辽宁 沈阳 110034)

**摘 要:** 潜在语义索引被广泛应用于信息检索、文本分类、自动问答等领域中。潜在语义索引是一种降维方法,它把共现特征映射到同一维空间上,而非共现特征映射到不同的空间上。在潜在语义索引的语义空间中,共现特征通过文档内部以及文档之间的特征传递关系获得。该文认为这种特征传递关系会引入一些不存在的共现特征,从而降低潜在语义索引的性能,应该对这种特征传递关系进行一些选择,削除不存在的共现特征信息。该文采用文档频率对文档集合进行特征选择,用 Complete-Link 聚类算法在两个公开语料上进行三个实验,实验结果显示,保留文档频度的 10%~15% 时,其 F1 值分别提高了 6.577 0%,1.992 8% 和 3.361 4%。

**关键词:** 计算机应用;中文信息处理;潜在语义索引;共现特征;奇异值分解;特征选择

**中图分类号:** TP391 **文献标识码:** A

## Research on Feature Optimization in Latent Semantic Indexing

JI Duo, ZHENG Wei, CAI Dong-feng

(Knowledge Engineering Center, Shenyang Institute of Aeronautical Engineering,  
Shenyang, Liaoning 110034, China)

**Abstract:** Latent Semantic Indexing (LSI) has been applied to many fields, such as information retrieval, text classification, automatic question answering and so on. Basically, LSI is a dimensionality reducing method by projecting term co-occurrences into the same space. Therefore, in the semantic space of LSI, term co-occurrences are obtained by the term transfer relation both in single document and between different documents. This paper suggests that this term transfer relation causes some nonexisted term co-occurrences, which reduce the performance of the LSI. To eliminate nonexisted term co-occurrences, this paper further adopts documents frequency to select features in document sets, and experiments with Complete-Link clustering algorithm on two public corpora. The experimental results show that the F-measure of clustering increases by 6.577 0%, 1.992 8% and 3.361 4% when documents frequency are reserved between 10% and 15%.

**Key words:** computer application; Chinese information processing; latent semantic indexing; term co-occurrence; singular value decomposition; feature selection

## 1 引言

潜在语义索引 (Latent Semantic Indexing, LSI) 在文本数据任务学习中取得了广泛的应用, LSI 应用于搜索引擎和信息检索、文本分类、信息过滤和文本聚类等领域。在国外, LSI 已经进入商业

化应用阶段,在国内,近年来这方面的研究也取得了许多进展,但是针对汉语的一些特点,尚有多个难点有待进一步解决。

潜在语义索引最初基于这样一个假设: 出现在文档中的词语并不是完全随机的,而是存在某种潜在语义结构。如果能把这种潜在语义结构提取出来,建立词语与词语之间的语义关系,就可以

收稿日期: 2008-08-30 定稿日期: 2008-10-02

基金项目: 国家 863 计划课题资助项目(2006AA01Z148);教育部科学技术研究重点项目(207148)

作者简介: 郑伟(1981—),男,硕士,主要研究方向为自然语言处理、文本聚类;季铎(1981—),男,硕士,主要研究方向为自然语言处理;蔡东风(1958—),男,博士,教授,主要研究方向为自然语言处理和人工智能。

削除词语用法的多样性和词语使用的随意性。然而,潜在语义索引并没有获得词语的语义信息,它仅仅捕获了词语的共现信息,词语的共现信息是通过词语之间的传递性建立的<sup>[9]</sup>。在一个文档集合中,文档中词语之间的隐含共现信息和文档集合中文档与文档之间词语的隐含共现信息是利用奇异值分解(Singular Value Decomposition, SVD)生成的。

文档中词语之间传递关系的多少对 SVD 生成词语共现信息有一定的影响,即词语之间传递关系越多,词语之间会产生一些不存在的共现信息,应该对词语之间的传递关系进行一些选择,本文认为词语之间传递关系的选择,可以通过对文档集合进行特征选择来完成。首先用文档频率(Document Frequency, DF)的方法进行特征选择,然后利用 SVD 分解得到词语共现信息,并用 Complete-Link 聚类算法在两个语料上进行了三个实验,结果显示保留文档频度的 10%~15%,选择合理的 K 值,聚类性能有明显提高。

## 2 潜在语义索引

潜在语义索引以奇异值分解(SVD)技术为基础。奇异值分解将特征文档矩阵  $A_{mn}$  分解为三个矩阵,如公式(1):一个特征维度矩阵  $T$ ,一个奇异值矩阵  $S$  和一个文档维度矩阵  $D$ 。

$$A_{mn} = TSD^T \quad (1)$$

其中,  $T$  和  $D$  为正交矩阵,  $S$  为对角矩阵,其对角线上的值由大到小排列。然后保留  $T$  和  $D$  中前  $K$  个列向量和  $S$  中的前  $K$  个奇异值,分别得到  $T_K$ 、 $D_K$  和  $S_K$ ,如公式(2)

$$A_K = T_K S_K D_K^T \quad (2)$$

$A_K$  是原始矩阵  $A_{mn}$  在秩为  $K$  条件下的最小二乘意义上的最优近似,上述过程被称为截断的奇异值分解,如图 1 所示。 $T$  保留了前  $K$  列,  $D^T$  保留了前  $K$  行,  $S$  保留了前  $K$  行和前  $K$  列。 $A_K$  的每一行代表一个特征,那么特征与特征之间的相似度如公式(3),因此,  $T_K S_K$  的行向量分别代表词语向量。 $A_K$  的每一列代表一篇文档,文档与文档之间的相似度如公式(4),因此,  $D_K S_K$  的行向量分别代表文档向量。

$$\begin{aligned} A_K A_K^T &= T_K S_K D_K^T (T_K S_K D_K^T)^T \\ &= T_K S_K S_K T_K^T \\ &= T_K S_K (T_K S_K)^T \end{aligned} \quad (3)$$

$$\begin{aligned} A_K^T A_K &= (T_K S_K D_K^T)^T T_K S_K D_K^T \\ &= D_K S_K^T T_K^T T_K S_K D_K T_K^T \\ &= D_K S_K (D_K S_K)^T \end{aligned} \quad (4)$$

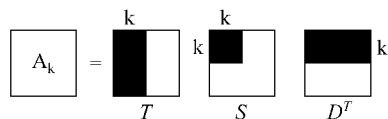


图 1 截断的奇异值分解

潜在语义索引的数学基础 SVD 第一次在文献[1]中论述并且由文献[2-3]做了进一步讨论。这些论文论述了 SVD 处理过程,并用几何内容解释了 SVD 分解后矩阵的意义。截断的 SVD 分解是原始矩阵的最优近似,文献[4]证明了 LSI 的性能主要来自 SVD 分解技术。其他研究者提出了一些理论方法来理解 LSI。Zha and Simon 用特征子空间模型来描述 LSI 并提出了用统计方法选择截断奇异值分解的最优维度<sup>[5]</sup>。Ding 用余弦相似度的方法为 LSI 构建了一个统计模型<sup>[6]</sup>,反映了特征和文档矩阵最大可能的形成,LSI 对于这个模型是最佳的解决方案。文献[9]认为 LSI 仅仅获得了特征的共现信息,并用数学的方法进行了证明,本文在文献[9]的基础上,认为特征之间的共现信息可以通过对 SVD 分解前的文档集合进行特征选择,从而减少在 SVD 空间中产生的不存在的特征共现信息,使文档在降低维度的基础之上能更近似地表示文档的内容,提高 LSI 的性能。

## 3 特征传递关系的选择

文献[9]认为 SVD 分解后能够捕获特征之间的共现信息。文献[1]中的例子如表 1 所示,该例子未进行 SVD 分解的特征文档矩阵如表 2 所示,表 3 是 SVD 分解前特征与特征的相似度矩阵,由公式(3)得知,分解后的相似度矩阵如表 4,与分解前的相似度矩阵表 3 相比,表 4 中特征与特征相似度的权值发生了明显的变化。在表 3 中,许多特征之间的相似度为 0,而 SVD 分解后表 4 中不存在相似度为 0 的值,即有些特征之间的共现信息被提高,而有些特征之间的共现信息被减弱。分解前有些特征与特征之间的相似度为 0,即特征之间没有联系或联系很小,在潜在语义空间中,这些特征与特征之间的相似度被加强。以 *user(t4)* 和 *human(t1)* 为例,分解前其相似度为 0,分解后相似度为 0.955 4。相似度值的变化可认为 *user* 与 *interface* 共现, *interface* 和

human 共现,因此,user 和 human 共现,在 LSI 的空间 user 和 human 被投影到同一维空间。

表 1 技术备忘录标题

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

表 2 Deerwester 特征文档矩阵

	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

特征之间的相似度反映了特征与特征的相关性,在 SVD 生成的空间中,其权值大小不仅体现了特征之间相关性,而且体现了特征与特征之间共现的信息,若特征的传递次数过多,特征的共现信息就越多,会产生过多不存在的共现信息,特征之间的相

表 3 Deerwester 原始的特征与特征相似度矩阵

	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12
t1	2	1	1	0	2	0	0	1	0	0	0	0
t2	1	2	1	1	1	0	0	1	0	0	0	0
t3	1	1	2	1	1	1	1	0	1	0	0	0
t4	0	1	1	3	2	2	2	1	1	0	0	0
t5	2	1	1	2	6	1	1	3	1	0	0	0
t6	0	0	1	2	1	2	2	0	1	0	0	0
t7	0	0	1	2	1	2	2	0	1	0	0	0
t8	1	1	0	1	3	0	0	2	0	0	0	0
t9	0	0	1	1	1	1	1	0	2	0	1	1
t10	0	0	0	0	0	0	0	0	0	3	2	1
t11	0	0	0	0	0	0	0	0	1	2	3	2
t12	0	0	0	0	0	0	0	0	1	1	2	2

似度权值将发生变化。这种特征共现信息不仅通过一个文档中特征与特征之间的传递性体现,而且也体现在文档与文档之间特征的传递性上。随着文档中和文档之间特征传递次数的增多,特征的共现信息会增加。但有些特征共现信息是不存在的,这就给文档集合中增加了一些噪音数据。从表 4 可以看出,time(t7)和 graph(t11)的相似度为 0.538 0,而这两个特征分别来自不同的类别,特征之间相似度的变化可以认为 time 与 user 共现,user 与 graph 共现,得到 time 与 graph 共现。

假设在这九篇文章中,每个文档中存在一个共同的特征,在该文档集合生成的语义空间中,由于特征之间相互传递,文档中会出现一些不存在的特征共现信息,从而使得文档与文档之间也产生一些不存在的特征共现信息,这些特征共现信息是噪音数据。在表 1 的每个文档中增加一个不存在的特征 X,其特征权值都为 1。将表 1 的特征文档矩阵进行 SVD 分解,用公式(3)计算特征与特征之间的相似度,其相似度矩阵如表 5。从表 5 中可以看出,computer(t3)与 response(t6)和 time(t7)权值均为0.692 5,与表 4 中的对应项相比,其权值被减弱,从表 1 中可以看出,这些词的权值应该非常相近,但由于引进了一个共有特征 X,其权值被减弱。

表 4 Deerwester 截断到二维的特征与特征相似度矩阵

	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12
t1	0.626 0	0.541 0	0.562 6	0.955 4	1.714 6	0.576 5	0.576 5	0.846 5	0.308 7	−0.326 9	−0.366 1	−0.250 9
t2	0.541 0	0.469 6	0.510 4	0.863 8	1.499 7	0.534 8	0.534 8	0.729 4	0.326 8	−0.200 2	−0.210 3	−0.139 7
t3	0.562 6	0.510 4	0.657 5	1.099 2	1.683 2	0.741 2	0.741 2	0.768 0	0.629 0	0.170 9	0.270 6	0.210 8
t4	0.955 4	0.863 8	1.099 2	1.839 2	2.841 6	1.233 4	1.233 4	1.303 1	1.028 5	0.238 1	0.392 4	0.309 2
t5	1.714 6	1.499 7	1.683 2	2.841 6	4.816 8	1.790 7	1.790 7	2.316 7	1.185 5	−0.438 1	−0.413 2	−0.258 5
t6	0.576 5	0.534 8	0.741 2	1.233 4	1.790 7	0.858 2	0.858 2	0.792 1	0.798 6	0.377 0	0.538 0	0.405 8
t7	0.576 5	0.534 8	0.741 2	1.233 4	1.790 7	0.858 2	0.858 2	0.792 1	0.798 6	0.377 0	0.538 0	0.405 8
t8	0.846 5	0.729 4	0.768 0	1.303 1	2.316 7	0.792 1	0.792 1	1.139 0	0.441 7	−0.404 5	−0.447 0	−0.304 5
t9	0.308 7	0.326 8	0.629 0	1.028 5	1.185 5	0.798 6	0.798 6	0.441 7	0.957 0	0.895 8	1.184 0	0.869 4
t10	−0.326 9	−0.200 2	0.170 9	0.238 1	−0.438 1	0.377 0	0.377 0	−0.404 5	0.895 8	1.553 9	1.977 2	1.431 1
t11	−0.366 1	−0.210 3	0.270 6	0.392 4	−0.413 2	0.538 0	0.538 0	−0.447 0	1.184 0	1.977 2	2.520 3	1.825 4
t12	−0.250 9	−0.139 7	0.210 8	0.309 2	−0.258 5	0.405 8	0.405 8	−0.304 5	0.869 4	1.431 1	1.825 4	1.322 4

表 5 增加特征后截断到二维的特征与特征相似度矩阵

	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13
T1	0.669 8	0.605 8	0.654 0	1.082 3	1.754 2	0.687 8	0.687 8	0.853 0	0.461 9	−0.009 2	−0.027 0	−0.019 0	1.762 1
T2	0.605 8	0.552 9	0.601 7	0.986 8	1.572 8	0.632 3	0.632 3	0.762 0	0.456 2	0.092 3	0.089 5	0.061 5	1.737 7
T3	0.654 0	0.601 7	0.659 5	1.073 1	1.684 7	0.692 5	0.692 5	0.813 5	0.529 6	0.196 6	0.206 3	0.142 6	2.014 8
T4	1.082 3	0.986 8	1.073 1	1.761 5	2.812 4	1.127 7	1.127 7	1.363 1	0.807 8	0.146 0	0.138 5	0.095 1	3.077 5
T5	1.754 2	1.572 8	1.684 7	2.812 4	4.631 2	1.773 3	1.773 3	2.259 6	1.105 1	−0.297 6	−0.380 3	−0.265 8	4.223 0
T6	0.687 8	0.632 3	0.692 5	1.127 7	1.773 3	0.727 2	0.727 2	0.856 5	0.552 7	0.195 8	0.204 6	0.141 3	2.103 2
T7	0.687 8	0.632 3	0.692 5	1.127 7	1.773 3	0.727 2	0.727 2	0.856 5	0.552 7	0.195 8	0.204 6	0.141 3	2.103 2
T8	0.853 0	0.762 0	0.813 5	1.363 1	2.259 6	0.856 5	0.856 5	1.104 1	0.515 8	−0.200	−0.248 6	−0.173 4	1.973 0
T9	0.461	0.456 2	0.529 6	0.807 8	1.105 1	0.552 7	0.552 7	0.515 8	0.612 4	0.762 4	0.851 6	0.590 7	2.316 0
t10	−0.009	0.092 3	0.196 6	0.146 0	−0.297	0.195 8	0.195 8	−0.200	0.762 4	2.010 8	2.276 5	1.580 3	2.855 4
t11	−0.027	0.089 5	0.206 3	0.138 5	−0.380	0.204 6	0.204 6	−0.248 6	0.851 6	2.276 5	2.577 7	1.789 4	3.188 7
t12	−0.019	0.061 5	0.142 6	0.095	−0.265	0.141 3	0.141 3	−0.173 4	0.590 7	1.580 3	1.789 4	1.242 2	2.211 9
t13	1.762 1	1.737 7	2.014 8	3.077 5	4.223 0	2.103 2	2.103 2	1.973 0	2.316 0	2.855 4	3.188 7	2.211 9	8.759 6

计算文档与文档之间的相似度时,文档之间的相似程度主要依赖于文档之间共现特征的多少,在生成的潜在语义空间中,由于特征之间的传递性,特征之间的潜在联系被挖掘出来。原来相似度很小或没有任何相同特征的文档,可能存在很高的相似性。文档与文档之间相似度计算如公式(4)所示,表 6 是未增加特征 X 时,文档与文档之间的相似度矩阵。从该矩阵的数据可以看出,文档之间具有明显的区

分性,同类文档具有较高的相似性,不同类文档具有较小的相似性。*m4* 和 *c2* 具有较高相似性,其相似度为 1.121 3。由于这两个文档中都有特征 *survey*,与 *survey* 共现的特征的权值被加强。同样若每个文档中有一个共有特征 X,其文档与文档之间的相似度矩阵如表 7。从表 7 中可以看出,文档之间的相似度不如表 6 中那么明显。

表 6 文档与文档相似度矩阵

	c1	c2	c3	c4	c5	m1	m2	m3	m4
c1	0.455 1	1.275 3	1.065 9	1.278 1	0.577 2	−0.061 3	−0.125 9	−0.169 0	−0.010 9
c2	1.275 3	4.275 9	2.994 9	3.418 8	2.004 5	0.232 1	0.567 4	0.821 3	1.121 3
c3	1.065 9	2.994 9	2.496 5	2.991 6	1.356 2	−0.138 9	−0.284 5	−0.381 2	−0.012 4
c4	1.278 1	3.418 8	2.991 6	3.627 2	1.531 2	−0.265 6	−0.567 1	−0.774 9	−0.297 5
c5	0.577 2	2.004 5	1.356 2	1.531 2	0.945 4	0.144 9	0.347 7	0.499 6	0.621 2
m1	−0.061 3	0.232 1	−0.138 9	−0.265 6	0.144 9	0.240 4	0.546 1	0.767 4	0.663 7
m2	−0.125 9	0.567 4	−0.284 5	−0.567 1	0.347 7	0.546 1	1.241 0	1.744 0	1.512 5
m3	−0.169 0	0.821 3	−0.381 2	−0.774 9	0.499 6	0.767 4	1.744 0	2.450 9	2.128 0
m4	−0.010 9	1.121 3	−0.012 4	−0.297 5	0.621 2	0.663 7	1.512 5	2.128 0	1.889 2

表 7 增加特征后文档与文档的相似度矩阵

	c1	c2	c3	c4	c5	m1	m2	m3	m4
c1	1.394 2	2.535 8	2.103 9	2.297 5	1.579 9	0.905 4	1.061 0	1.168 7	1.227 8
c2	2.535 8	4.810 8	4.083 9	4.554 1	2.892 9	1.306 0	1.356 8	1.391 6	1.665 5
c3	2.103 9	4.083 9	3.508 2	3.953 4	2.409 1	0.924 5	0.858 5	0.812 2	1.117 0
c4	2.297 5	4.554 1	3.953 4	4.495 8	2.640 0	0.847 5	0.664 9	0.537 8	0.949 7
c5	1.579 9	2.892 9	2.409 1	2.640 0	1.792 3	0.993 0	1.146 8	1.253 2	1.336 3
m1	0.905 4	1.306 0	0.924 5	0.847 5	0.993 0	1.173 5	1.673 0	2.019 8	1.772 4
m2	1.061 0	1.356 8	0.858 5	0.664 9	1.146 8	1.673 0	2.461 2	3.008 4	2.573 2
m3	1.168 7	1.391 6	0.812 2	0.537 8	1.253 2	2.019 8	3.008 4	3.694 8	3.129 0
m4	1.227 8	1.665 5	1.117 0	0.949 7	1.336 3	1.772 4	2.573 2	3.129 0	2.705 2

在 LSI 的空间,特征与特征之间的传递性,使得文档与文档之间的相似度发生变化。从特征之间相似度的变化和文档之间相似度的变化可以看出,共现特征会产生一些不存在的共现信息,产生一些噪音数据。应该对特征与特征之间的传递关系进行一些选择,过滤掉不存在的特征传递。

在过滤不存在的特征传递时,在 LSI 生成的空间,特征与特征之间的传递关系已经形成,因此,不能在该语义空间进行特征选择。应该在 LSI 处理以前,对文档集合进行特征选择。本文采用文档频率 DF(Document Frequency)<sup>[10]</sup>对文档集合进行特征选择,DF 不仅考虑文档中特征之间传递的关系,而且考虑了文档与文档之间特征之间传递的关系。本文首先对文档集合进行 DF 特征过滤,以减少 LSI 空间不存在的特征共现信息。然后对生成的特征文档矩阵进行 SVD 分解,在潜在语义空间选择合

理的 K 值,形成新的特征空间。

4 实验

4.1 语料

本实验所采用的语料来源于中国科学院谭松波博士提供的语料 Tancorpv1.0 以及搜狗实验室文本分类语料。从 Tancorpv1.0 中的 12 个大类中选取 12 个类,总共 2 400 篇文本,记为中国科学院语料 1,文本最小的为 1KB,最大的文本 14.7KB,具体类别中文本数量的分布情况如表 8,从全部 60 个小类中选取 3 000 篇文本,记为中国科学院语料 2,具体分布情况如表 9。从搜狗实验室下载的文本分类语料库中选取 9 个类中的 1 000 篇文本,其中最大的类包含 200 篇文档,最小的类包含 80 篇文档,具体分布情况如表 10。

表 8 中国科学院语料 1

类别	数量	类别	数量
地域	112	体育	201
汽车	199	房产	201
财经	200	艺术	208
卫生	200	教育	205
娱乐	210	人才	215
科技	217	电脑	232

表 9 中国科学院语料 2

类别	数量	类别	数量	类别	数量	类别	数量
财富	16	篮球	141	羽球	29	培训	9
金融	49	乒乓	67	足球	160	校园	106
企业	86	棋牌	12	出版	19	招生	49
人物	50	水上	51	就业	78	保健	52
消费	51	田径	44	考试	76	两性	50
城建	65	自然科学		115	音乐艺术		49
私宅	92	电子商务		47	考古科学		92
装修	64	人才创业		15	音乐娱乐		65
组屋	47	汽车百科		15	汽车政策		20
心理	47	地域风俗		28	古董艺术		28
证券	59	天文科学		86	生命科学		116
网球	37	电影娱乐		70	电脑游戏		30
留学	40	电脑科技		27	汽车行驶		35
医药	27	综艺娱乐		76	汽车快讯		30
文学	106	电脑网络		21	人才薪金		26
电脑病毒		23	人才猎取		27		
地域城市		17	美学艺术		38		
电脑软件		14	舞台艺术		22		
地域美食		20	人才应试		25		
人才管理		12	人才履历		32		

表 10 搜狗实验室语料

类别	数量	类别	数量
IT	80	教育	100
军事	80	旅游	120
招聘	100	健康	120
体育	100	文化	200
财经	100		

4.2 评测方法

本文中聚类效果的评价参照了信息检索中的评测方法,将每个聚类结果看作是查询的结果,这样,对于最终的某一个聚类类别  $r$  和原来的预定类别  $i$ ,其  $F\text{-Measure}^{[11-12]}$  的准确率 ( $precision$ ) 和召回率 ( $recall$ ) 的定义,如下:

$$recall(i,r) = n(i,r)/n_i \tag{5}$$

$$precision(i,r) = n(i,r)/n_r \tag{6}$$

其中  $n(i,r)$  是聚类中包含类别  $i$  中的文本数目,  $n_r$  是聚类形成的类别  $r$  包含的文本数目,  $n_i$  是预定义类别的  $i$  包含的文本数目。于是,聚类  $r$  和类别  $i$  之间的  $F\text{-Measure}$  值计算如下:

$$F(i,r) = \frac{2 \times recall(i,r) \times precision(i,r)}{precision(i,r) + recall(i,r)} \tag{7}$$

每个类的  $F\text{-Measure}$  是该类别在所有类中获得的最大的值,根据每个类的  $F$  值  $F\text{-Measure}$ ,聚类结果性能的最终评测方法用  $MacroF1$ ,其定义如下:

$$MacroF1 = \sum_i \frac{n_i}{n} \max\{F(i,r)\} \tag{8}$$

4.3 实验结果与分析

本文首先对实验语料进行特征选择,针对不同的实验语料设置不同的阈值  $\alpha \times FT$  ( $FT$ ,各个实验语料的文档总数; $\alpha$ ,比例因子其取值范围在  $(0,1]$ ;  $\alpha \times FT$  的值取整数),  $DF_{ij}$  是第  $i$  篇文档第  $j$  个特征的文档频率,将  $DF_{ij} > \alpha \times FT$  的特征过滤掉,形成新的特征空间,特征的权重采用  $TF\text{-IDF}^{[18]}$  计算,用向量空间模型生成过滤后的特征文档矩阵。然后,用  $SVD$  对该特征文档矩阵进行分解<sup>[14]</sup>。在  $LSI$  空间,文本的相似性采用向量夹角余弦的计算方法,聚类处理阶段采用  $Complete\text{-Link}$  算法。

考察将文档频率  $DF_{ij} > \alpha \times FT\%$  的特征过滤时,不同  $\alpha$  所对应的特征选取方案,在不同  $SVD$  空间对聚类性能的影响是不同的。各语料的实验结果如下面 3 个表所示,从这些表中可以看出,随着  $\alpha$  的减少,最优  $K$  值所对应的  $Macro\text{-}F$  值都呈现出先升后降的趋势。这表明选取合适的  $\alpha$ ,对文档集合的特征进行合理的选择,不仅可以减少文档中特征与特征之间的共现信息,还可以减少文档与文档之间特征的共现信息,消除  $SVD$  空间中的数据噪音。在降低文档特征空间的同时,可以更近似地表示文本的语义信息,从而提高聚类算法的速度和性能。

$\alpha=100\%$  时,是未进行特征选择的情况,从三个

语料的实验结果可以看出,经过特征选择并且在 LSI 空间选择合理的 K 值,其聚类性能比未进行特征选择的聚类性能有明显的提高。在搜狗语料上实验结果见表 11,  $DF > 85\% FT$ ,  $K = 10$  时,其  $Macro-F$  值达到了  $65.3312\%$ ,比未过滤特征的  $Macro-F$  值提高了大约  $6.5770\%$ ;在中国科学院语

料 1 实验结果见表 12,  $DF > 87\% FT$ ,  $K = 10$  时,其  $Macro-F$  值为  $75.8125\%$ ,比未过滤特征的  $Macro-F$  值提高了约  $1.9928\%$ ;而在中国科学院语料 2 实验结果见表 13,  $DF > 88\% FT$ ,  $K = 110$  时,其  $Macro-F$  值为  $58.9859\%$ ,比未过滤特征的  $Macro-F$  值提高了  $3.3614\%$ 。

表 11 搜狗实验室语料实验结果

<div><div>K</div><div><math>\alpha</math></div></div>	100%	95%	90%	85%	70%	50%	30%
c5	51.828 9	45.121 7	52.943 5	47.673 9	51.401 6	50.070 3	50.070 3
10	58.754 2	57.033	58.67	65.331 2	62.531 2	58.754 2	58.754 2
30	54.583 7	53.177 6	49.910 2	55.47	47.200 7	54.583 7	54.583 7
50	47.742 9	48.539 9	41.737 8	46.738 7	37.025 5	47.742 9	47.742 9
100	36.888 5	33.434 7	32.576 8	32.754 1	36.056 2	36.888 5	36.888 5
150	37.328 7	37.408 7	35.139 8	38.515 1	39.270 8	37.328 7	37.328 7
200	29.183	34.033 1	33.822 1	35.457 4	32.577	29.183	29.183

表 12 中国科学院语料 1 实验结果

<div><div>K</div><div><math>\alpha</math></div></div>	100%	95%	90%	87%	70%	50%	30%
5	45.604	57.73	58.626 7	57.659	56.536 3	54.581 7	54.581 7
10	73.819 7	67.685 5	71.577 2	75.812 5	71.297 4	73.019 7	73.019 7
30	52.593 5	53.543 4	53.172 3	51.696 9	49.863 9	52.593 5	52.593 5
50	46.378 4	43.130 1	45.259	43.105 1	40.244	46.378 4	46.378 4
100	48.596 5	38.854 5	38.750 9	36.027 1	42.918 8	48.596 5	48.596 5
150	34.740 2	32.921 3	32.651 2	38.217 4	45.364 5	34.740 2	34.740 2
200	38.942 4	29.089 7	35.548 8	34.138	34.306	38.942 4	38.942 4

表 13 中国科学院语料 2 实验结果

<div><div>K</div><div><math>\alpha</math></div></div>	100%	90%	88%	70%	50%	30%
5	30.485	29.309	28.001 3	29.130 4	26.334 3	26.334 3
10	41.628 4	39.222 6	41.157	39.454 8	37.942 6	37.942 6
30	51.412 6	54.537 5	54.496 5	56.224 8	55.687 8	55.687 8
50	55.624 5	55.537 5	54.946 2	55.188 5	55.921 8	55.921 8
100	49.709 8	56.323 5	56.664 1	52.183 7	54.852 6	54.852 6
110	52.998 3	55.891 1	58.985 9	56.854 7	57.390 1	57.390 1
150	50.568	53.529 7	54.908 7	49.099 4	50.868 1	50.868 1
200	49.483 4	46.696	51.368 8	50.509 7	48.511 3	48.511 3

从各个语料的实验数据可以看出,过滤较多特征时,将会过滤一些在特征与特征之间起重要传递作用的特征,在 SVD 空间将不会产生应有的特征共现信息,本来存在的共现特征不会投影到同一维空

间上,其特征与特征的相似性减弱,从而文档与文档之间的相似性也将减弱。相反,过滤较少特征时,不存在的特征共现信息将会保留,在 SVD 空间这些共现特征被投影到同一维空间上,特征与特征的相似

性增加,其文档与文档之间的相似性也将变大。在搜狗实验室语料和中国科学院语料 1 上,将文档频度小于 5%FT 的特征分别作为新的特征集合时,其最优  $K$  值对应的  $Macro-F$  值均小于未过滤特征时的  $Macro-F$  值。若过滤较少特征时,不能过滤掉在特征与特征之间产生噪音的特征,当各个语料将文档频度小于 50%FT 的特征作为新的特征集合时,其  $Macro-F$  值将不再变化。当保留特征文档频度的 10%~15% 时,能有效过滤掉一些噪音数据。所以,针对各个语料过滤掉合理的特征时,文档频度的方法能够过滤掉一些特征与特征之间传递的次数,在 SVD 空间中减少特征与特征之间的传递次数,减少不存在共现特征,使文档与文档之间的区分能力更明显,削除噪音数据,聚类的  $Macro-F$  值有明显的提高。

## 5 结论与展望

本文认为在 SVD 空间中,特征与特征之间的传递次数对潜在语义索引的性能有很大的影响。若特征之间传递次数过多,就会产生一些不存在的特征共现信息,影响特征与特征之间的相似度,使相似性很小的文档之间的相似度变大,从而影响潜在语义索引的性能。在 SVD 分解前对文档集合中的特征进行 DF 特征选择,可以降低特征与特征之间的传递次数,从而减少不存在的特征共现信息,使文档之间的相似度更加明显。实验结果显示该方法能够提高文本聚类的性能。本文采用 DF 方法对文档集合中文档进行特征选择,对特征与特征之间的传递次数做了简单的过滤。下一步的工作是基于文档中特征与特征的条件熵和文档之间的条件熵进行特征选择。

## 参考文献:

- [1] Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., Harshman, R. A. Indexing by latent semantic analysis [J]. Journal of the American Society of Information Science, 1990, 41(6), 391-407.
- [2] Berry, M. W., Dumais, S. T., O'Brien, G. W. Using linear algebra for intelligent information retrieval [J]. SIAM Review, 1995, 37(4), 575-595.
- [3] Berry, M. W., Drmac, Z., Jessup, E. R. Matrices, vector spaces, and information retrieval [J]. SIAM Review, 1999, 41(2), 335-362.
- [4] Wiener-Hastings, P. How latent is latent semantic analysis? [C]//Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, 1999: 932-937.
- [5] Ding C. H. Q. A similarity-based probability model for latent semantic indexing [C]//Proceedings of the Twenty-second Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, 1999: 59-65.
- [6] A. Kontostathis and W. M. Pottenger. A framework for understanding Latent Semantic Indexing (LSI) performance [J]. Information Processing and Management, 2006, 42(1), 56-73.
- [7] A. Kontostathis, W. M. Pottenger, and B. D. Davidson. Identification of critical values in latent semantic indexing [M]. T. Lin, S. Ohsuga, C. Liau, X. Hu, and S. Tsumoto editors, Foundations of Data Mining and Knowledge Discovery [M], Springer-Verlag, 2005: 333-346.
- [8] A. Kontostathis, L. E. Holzman, and W. M. Pottenger. Use of term clusters for emerging trend detection [Z]. Preprint, 2004.
- [9] 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.
- [10] Yiming Yang, Jan O. A Comparative Study on Feature Selection in Text Categorization [C]//Pedersen Proceedings of ICML-97, 14th International Conference on Machine Learning table of contents, 1997: 412-420.
- [11] Ying Zhao, George Karypis. Evaluation of Hierarchical Clustering Algorithms for Document Datasets [C]//Proceedings of the International Conference on Information and knowledge Management, New York, 2002: 515-524.
- [12] Michael Steinbach et al. A Comparison of Document Clustering techniques [R]. Technical report of University of Minnesota, KDD2000.
- [13] 刘云峰,齐欢等.潜在语义分析权重计算的改进 [J]. 中文信息学报, 2005, 19(6): 64-69.
- [14] PD Dr. Karin Haenelt and Prof. Dr. Peter Hellwig. Latent Semantic Indexing and Information Retrieval A quest with BosSE [R]. Seminar für Computerlinguistik Institut für allgemeine und angewandte Sprachwissenschaft Ruprecht-Karls-Universität Heidelberg Magisterarbeit 18 January 2006.
- [15] 林鸿飞,战学刚,姚天顺.文本层次分析与文本浏览 [J]. 中文信息学报, 1999, 14(5): 49-56.