基于 LDA 模型的文本相似度研究

陈 攀[1],杨 浩[1],吕 品[1,2],王海晖[1,2]
(1. 武汉工程大学 计算机科学与工程学院,湖北 武汉 430073;
2. 武汉工程大学 智能机器人湖北省重点实验室,湖北 武汉 430073)

摘 要:LDA 主题模型是近年来提出的一种具有文本表示能力的非监督学习模型。考虑到传统主题模型在处理大规模文本时存在的局限性,文中提出一种基于 LDA 模型的文本相似度计算方法。利用 LDA 为语料库建模,通过 Gibbs 抽样间接估算模型参数,将文本表示为固定隐含主题集上的概率分布,以此计算文本之间的相似度。最后将 K-means 算法作为文本相似度的评估指标。实验结果表明,与 LSI 模型相比,该方法能有效地提高文本相似度计算的准确性和文本聚类效果。

关键词:文本挖掘;LDA 模型;Gibbs 抽样;文本相似度
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2016)04-0082-04
doi:10.3969/j.issn.1673-629X.2016.04.18

Study on Text Similarity Based on LDA Model

CHEN Pan[1], YANG Hao[1], LÜ Pin[1,2], WANG Hai-hui[1,2]
(1. School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430073, China;
2. Hubei Province Key Laboratory of Intelligent Robot, Wuhan Institute of Technology, Wuhan 430073, China)

Abstract: LDA topic model is an unsupervised model which exhibits superiority on latent topic modeling of text data in the research of recent years. Considering the disadvantage of the traditional topic model when dealing with the large-scale text corpuses, a method which improves text similarity computations by using LDA model is proposed. It models corpus with LDA, parameters are estimated with Gibbs sampling. Each document is represented for the probability distribution of fixed implied theme set and computed the similarity between the texts. Finally, the K-means algorithm is selected as the evaluation index of text similarity. Experimental results show this method can improve the accuracy of text similarity and clustering quality of text effectively compared with LSI model.

Key words: text mining; LDA model; Gibbs sampling; text similarity

0 引 言

近年来,互联网作为一个开放的信息平台得到快速发展,网络上文本信息量也以指数级的方式飞速增长。在大数据时代,信息中包含很多数据,这些数据大部分以文本的形式存在。面对如此多的文本信息,如何高效地进行文本挖掘是目前研究的重点问题,这使得文本挖掘成为大数据时代信息处理领域的热点。

常用的文本挖掘方法是潜在语义索引 LSI(Latent Semantic Indexing)模型[1]。利用 LSI 模型对文本进行挖掘时,由于考虑了词间的语义关系,具有很好的降维效果,但对重要稀有类别的分类特征,LSI 模型可能过滤了它们,从而造成分类性能不佳。LDA(Latent Dirichlet Allocation)模型[2]改进了 LSI 模型在文本挖掘中的不足,有效解决了文本挖掘中的特征稀疏和分类性能受损问题。文中基于 LDA 模型进行文本相似度计算,采用传统的聚类算法对实验结果进行评估,并获得了较好的效果。

1 相关工作

在文本挖掘领域,国内外研究人员都进行了大量的工作。Salton 等提出向量空间模型(Vector Space Model,VSM)[3]是常用算法。Hastie 等提出 KNN(K-Nearest Neighbor)[4]方法来计算文本相似度。Blei 等[2]提出 LDA 主题模型。该模型以文本特征为对象,

收稿日期:2015-07-16 **修回日期**:2015-10-21 **网络出版时间**:2016-03-22
基金项目:湖北省高等学校优秀中青年团队计划项目(T201206);湖北省智能机器人重点实验室开放基金(HBIR201409)
作者简介:陈 攀(1993-),男,研究方向为文本挖掘与自然语言处理;吕 品,博士,副教授,研究方向为数据挖掘、情感分析;王海晖,博士,教授,硕士生导师,研究方向为智能系统与机器视觉。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20160322.1521.080.html

将文本语料表示为各个主题空间,通过找到文本中不同隐含主题与词间的关系,得到文本主题概率分布。

目前,国内研究人员主要对 LDA 算法进行改进。刘振鹿等[5]基于 LDA 模型研究潜在语义分析,将语义划分为三个不同频段的语义区。通过语义互作用机制和文本类别对聚类结果进行修正,得到了较好效果。李文波等[6]在传统主题模型中融入文本类别信息,提出了一种附加类别的 LDA 模型方法来提高 LDA 模型的分类性能。石晶等[7]基于 LDA 模型进行文本建模,结合特征词相关扩充和背景特征词聚类,把特征词应用到待分析的文本中,找到特征词下的文本语义,提高文本分析的性能。

2　LDA 模型

LDA 模型是由 Beli 提出的针对离散数据集[8]建模的主题生成模型,它是一个三层贝叶斯网络结构,分为文档层、主题层和词层。其有向概率图[9]如图 1 所示。

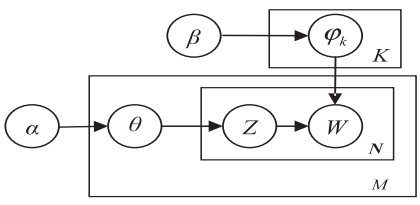

图 1　LDA 有向概率图

图 1 中,参数 α 反映出文本集中不同隐含主题间的相对强弱关系,参数 β 则代表主题自身的概率分布。Z 表示隐含主题, W 表示词表的每个词,即观察值。θ 代表文本-主题概率分布, φ_k 代表主题-词概率分布。对于给定的文本集 D ,包含 M 个文档, T 个主题,而每个文档 d 中又包含 N 个词。

2.1　相关符号含义

LDA 模型中的对应符号[10]解释如下:

(1)文本数据是由词组成的集合单元。词表序列用 $\{1,2,\cdots,N\}$ 表示,用向量 w 表示词表中第 v 个词,对于每个词 $u \neq v$, $w_u \neq 0$, $w_v \neq 1$。

(2)文档是由 N 个词组成的序列。该序列用 $d = \{w_1, w_2, \cdots, w_n\}$ 表示, w_n 代表序列的第 n 个词。

(3)文档集由文档组成,每个文档集中包含若干文档,用 $D = \{d_1, d_2, \cdots, d_m\}$ 表示,其中 d_m 表示第几篇文档。

假设主题数为 K ,则文档 d 中的第 i 个词 w_i 的概率为:

$$P(w_i) = \sum_{j=1}^{T} P(w_i \mid z_i = j) P(z_i = j) \tag{1}$$

式中: z_i 是隐含主题变量,表示第 i 个词 w_i 属于 z_i 主题; $P(w_i \mid z_i = j)$ 表示词 w_i 对主题 j 的贡献概率; $P(z_i = j)$ 表示文档 d 对主题 j 的贡献概率。

词 w 在文本 d 中"出现"的概率可表示为:

$$p(w \mid d) = \sum_{j=1}^{T} \varphi_w^j \cdot \theta_j^d \tag{2}$$

式中: $\varphi_{w_i}^j = P(w_i \mid z_i = j)$ 表示第 j 个主题中,词表中 V 个词的多项式概率分布; $\theta_j^d = P(z_i = j)$ 表示若干个隐含主题在文本集中的随机混合。

通过 EM 算法[11]可求出最大似然函数:

$$L(\alpha \mid \beta) = \sum_{i=1}^{M} \log P(d_i \mid \alpha, \beta) \tag{3}$$

其中, α , β 为最大似然估计量,通过估算 α 和 β 的参数值从而确定 LDA 模型。则文本 d "发生"的条件概率分布可用式(4)表示[12]:

$$P(d \mid \alpha, \beta) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \int \left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1}\right) \left(\sum_{n=1}^{N} \sum_{i=1}^{k} \sum_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j}\right) \mathrm{d}\theta \tag{4}$$

2.2　参数估计

在 MCMC[13]中 Gibbs 抽样[14]是间接计算 LDA 模型参数的常用有效方法。具体步骤如下:

(1)将主题 z_i 的值随机设定为 1 到 T 内某个整数, i 是语料库所有文本中特征词的个数,它与词表规模和所在位置有关。

(2)迭代足够多次,直到 Markov 链[15-16]接近目标分布,此时的主题 z_i 可按如下公式估算 φ 和 θ 的值:

$$\hat{\varphi}_j^{(w)} = \frac{n_j^{(w)} + \beta}{n_j^{(\mathrm{all})} + W\beta} \tag{5}$$

$$\hat{\theta}^{(d)} = \frac{n_j^{(d)} + \alpha}{n_{(\mathrm{all})}^{(d)} + T\alpha} \tag{6}$$

式中: $n_j^{(w)}$ 表示词 w 对主题 j 有贡献的词数; $n_j^{(\mathrm{all})}$ 表示所有对主题 j 有贡献的词数; $n_j^{(d)}$ 表示文本 d 对主题 j 有贡献的词数; $n_{(\mathrm{all})}^{(d)}$ 表示文本 d 的所有对主题 j 有贡献的词数。

LDA 模型容易处理语料之外的新文本,得到文本主题概率分布的方法是考虑词 w_i 对于主题的后验概率[17] $P(w \mid z)$ 。通过 Gibbs 抽样间接得到 φ 和 θ 的值,记为后验概率 $P(z_i = j \mid z_{-i}, w_i)$,其计算公式如下:

$$P(z_i = j \mid z_{-i}, w_i) \propto \frac{z_{-i,j}^{(w_i)} + \beta}{z_{-i,j}^{(\mathrm{all})} + W\beta} \cdot \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,(\mathrm{all})}^{(d_i)} + T\alpha} \tag{7}$$

式中: z_{-i} 表示所有主题 z_k $(k \neq i)$ 的概率分配; $z_{-i,j}^{(w_i)}$ 表示词 w_i 属于主题 j 的词数; $z_{-i,j}^{(\mathrm{all})}$ 表示分配给主题 j 的所有词数。在文本 d_i 中,属于主题 j 的词用 $n_{-i,j}^{(d_i)}$ 表示,所有属于主题 j 的词用 $n_{-i,(\mathrm{all})}^{(d_i)}$ 表示。

2.3 相似度计算

文本相似度计算的核心是通过计算文本间的主题概率分布来实现。当用 LDA 模型找到了文本的隐含主题后,文本的相似度可通过计算对应的隐含主题概率分布的相似度来表示。通常用 KL 距离[18]公式作为相似度度量的标准,公式如下:

$$D_{\mathrm{KL}}(p,q)=\sum_{j=1}^{T}p_i\cdot\ln\frac{p_i}{q_i} \tag{8}$$

对于所有 j,当 $p_j=q_j$ 时,$D_{\mathrm{KL}}(p,q)=0$。考虑到 KL 距离公式的不对称性,故常用其对称版本[19]:

$$D_\lambda(p,q)=\lambda D_{\mathrm{KL}}(p,\lambda p+(1-\lambda)q)+(1-\lambda)D_{\mathrm{KL}}(q,\lambda p+(1-\lambda)q) \tag{9}$$

当 $\lambda=0.5$ 时,KL 距离公式可转化为 JS 距离公式[20]:

$$D_{\mathrm{JS}}(p,q)=\frac{1}{2}\left[D_{\mathrm{KL}}\left(p,\frac{p+q}{2}\right)+D_{\mathrm{KL}}\left(q,\frac{p+q}{2}\right)\right] \tag{10}$$

文中采用 JS 距离公式作为文本相似度的度量标准,JS 距离的向量区间为 $[0,1]$。

3 实验设计和结果分析

文本相似度计算的步骤如图 2 所示。

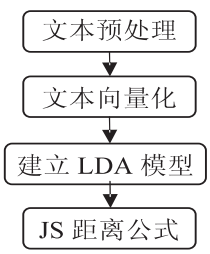

图 2 文本相似度计算

(1)对文本进行预处理,包括去除分词、停用词、符号等操作。

(2)将文本向量化,构成文本—词矩阵。

(3)利用向量化矩阵进行 LDA 建模,得到文本的主题概率分布。

(4)通过 JS 距离公式计算文本间的相似度,得到相似度矩阵。采用 K-means 算法对文本进行聚类,用聚类结果对文本相似度计算的准确性进行评估。

主题建模过程中,假定主题数 K 为 2,α 和 β 为经验值[21],分别为 $50/K$、0.1。为确保实验结果的准确性,Gibbs 抽样迭代次数需达到 1 000 次以上。改变主题数 K 值,根据聚类结果来评价最优主题数。

3.1 语料选择

实验预料数据来自复旦大学的一个英文语料库,共 6 个类别,2 246 篇文本。其中训练语料 50 篇,测试语料 2 196 篇,分别为 Science 类、Art 类、Business 类、Movie 类、Sport 类和 Travel 类。

3.2 评估方法

文本相似度计算的度量标准采用 JS 距离公式,最后利用 K-means 算法[22]对文本进行聚类。K-means 算法是一种较典型的逐点修改迭代的动态聚类[23]算法。采用信息检索中常用的一种平衡指标,K-means 算法中的 F 度量值[24]来衡量文本的相似度。F 度量值由查准率 $P(i,j)$ 和查全率 $R(i,j)$ [25]组成,查准率和查全率如式(11)[13]所示:

$$P(i,j)=\frac{n_{ij}}{n_j},R(i,j)=\frac{n_{ij}}{n_i} \tag{11}$$

式中:n_j 表示判断属于类别 j 的文本数目;n_i 表示实际属于类别 i 的文本数目;n_{ij} 表示判断属于 j 同时实际也属于 i 的文本数目。

F 度量值可定义[11]为:

$$F(i,j)=\frac{2\times P(i,j)\times R(i,j)}{P(i,j)+R(i,j)} \tag{12}$$

则文本集聚类的 F 度量值定义为:

$$F=\sum_i\frac{n_i}{n}\max_j(F(i,j)) \tag{13}$$

3.3 实验结果分析

当主题数 K 值为 2 时,α、β 分别为 $50/K$、0.1,测试文本为 50。此时迭代结果如图 3 所示。

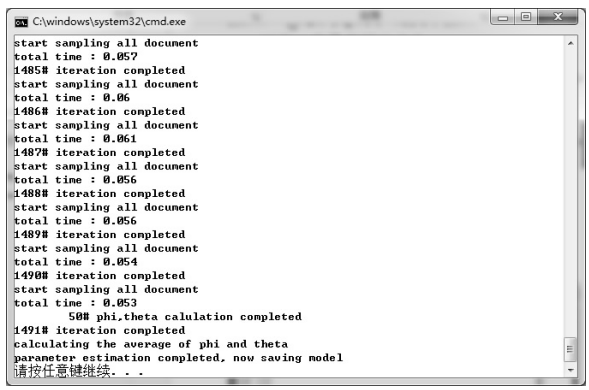

图 3 实验结果

改变主题数 K 的值,依次取值为 2、5、10、30、50、70、90。通过不同主题数进行多次聚类实验,确定最优主题数 K。

从图 4 中可以看出,当主题数 K 为 50 时,F 度量

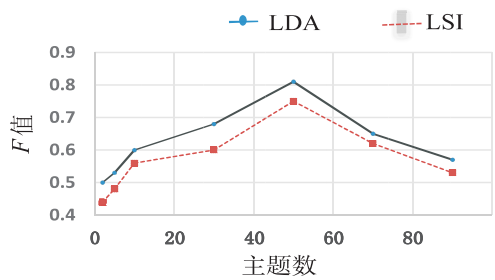

图 4 不同主题数的聚类效果

值最高,可以确定最优主题数为 50。同时,LDA 模型的聚类效果 F 度量值相比 LSI 模型更具有优势。

此外,主题数取值不同,实验的迭代时间会随主题数的增加而线性增长,如图 5 所示。

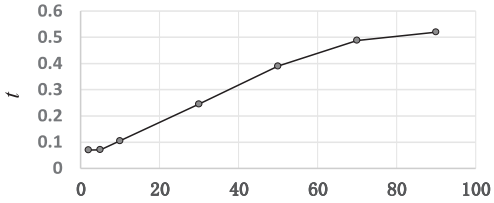

图 5　不同主题数下实验运行的时间

实验过程中,发现改变主题数 K 的值,相应的 α 值也会改变。K 值和 α 成反比关系,显然 K 值越小, α 值就越大,表明每个文档含更少的主题。β 一般为经验值,它表示每个主题分布在若干个词上。另外,训练语料的数目 S 会影响迭代次数 Ite ,二者成正比关系,但训练语料数目不会影响迭代时间 Ite - time 。

实验结果如表 1 所示。

表 1　主题数、训练语料数对实验的影响

α	β	K	S	Ite	Ite - time
25	0.1	2	50	1 491	0.050
10	0.1	5	50	1 491	0.071
5	0.1	10	50	1 491	0.105
1.7	0.1	30	50	1 491	0.245
1.0	0.1	50	50	1 491	0.390
0.71	0.1	70	50	1 491	0.488
0.56	0.1	90	50	1 491	0.521
25	0.1	2	30	1 291	0.050
25	0.1	2	10	1 091	0.050
25	0.1	2	5	1 041	0.050

4　LDA 应用于文本挖掘的研究展望

目前,基于 LDA 模型的主题句抽取方法应用广泛并取得了较好效果。下一步将重点研究如何选择大量未标注的可靠主题句来扩充训练 LDA 模型,以及如何使用关键词准确地抽取主题句以及候选主题句。通过两者相互促进,提高整体的抽取性能。

基于 LDA 模型的文本聚类比传统聚类效果更加优越,但这种方法只针对普通文档集。对于数字图书的特殊语料,则需要联合数字图书的信息目录和正文信息进行主题建模的方式进行聚类研究。

基于 LDA 模型的文本分割在预处理领域极为重要。实验研究过程中,除了需要直接测试,更需要间接测试,即将文本置于应用系统中考查,工作重点是进行更有效的测试。

5　结束语

文中介绍了 LDA 主题模型,该模型有效解决了 LSI 模型在文本挖掘中的特征稀疏和分类性能受损问题。实验结果表明,LDA 模型应用于文本相似度计算,相对于 LSI 模型更具有优越性,效率也更高。同时文中简要列举了 LDA 模型在文本挖掘中的不同应用,并总结了 LDA 模型在文本挖掘中面临的一些挑战和待解决的问题。LDA 模型应用于文本相似度计算,考虑到 LDA 模型具有易扩展性,下一步工作将在 LDA 模型的基础上,继续研究、改进文本建模方法及基于其上的文本挖掘。

参考文献:

[1] Deerwester S, Dumais S T A. Indexing by latent semantic analysis[J]. Journal of the Society for Information Science, 1990, 41(6):391-407.

[2] Blei D, Ng A, Jordan M. Latent Dirichlet allocation[J]. Journal of Machine Leaning Research, 2003, 3:993-1022.

[3] Salton G, Wong A, Yang C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11):613-620.

[4] Hastie T, Tibshirani R. Discriminant adaptive nearest neighbor classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996, 18(6):607-616.

[5] 刘振鹿,王大玲,冯　时,等. 一种基于 LDA 的潜在语义区划分及 Web 文档聚类算法[J]. 中文信息学报,2011,25(1):60-65.

[6] 李文波,孙　乐,张大鲵. 基于 Labeled-LDA 模型的文本分类新算法[J]. 计算机学报,2008,31(4):620-627.

[7] 石　晶,胡　明,石　鑫,等. 基于 LDA 模型的文本分割[J]. 计算机学报,2008,31(10):1865-1873.

[8] 徐　戈,王厚峰. 自然语言处理中主题模型的发展[J]. 计算机学报,2011,34(8):1423-1436.

[9] 王李冬,魏宝刚,袁　杰. 基于概率主题模型的文档聚类[J]. 电子学报,2012,40(11):2346-2350.

[10] 姚全珠,宋志理,彭　程. 基于 LDA 模型的文本分类研究[J]. 计算机工程与应用,2011,47(13):150-153.

[11] Andrzejewski D, Buttler D. Latent topic feedback for information retrieval[C]//Proceedings of 17th ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM Press, 2011:600-608.

[12] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers[J]. Machine Learning, 1997, 29(2-3):131-163.

[13] Doucet, Godsill S, Andrieu C. On sequential Monte Carlo sampling methods for Bayesian filtering[J]. Statistics and Computing, 2000, 10(3):197-208.

[14] 马海云. 基于 Gibbs 抽样的测试用例生成技术研究[J]. 自动化与仪器仪表,2011(3):11-12.

(下转第 89 页)

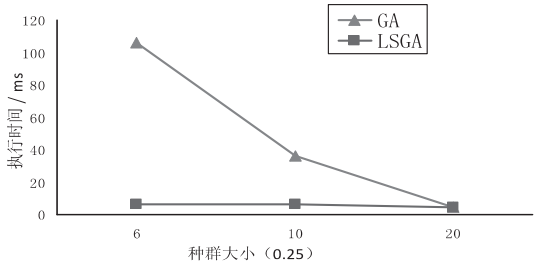

图 1　实验 1 运行时间对比图

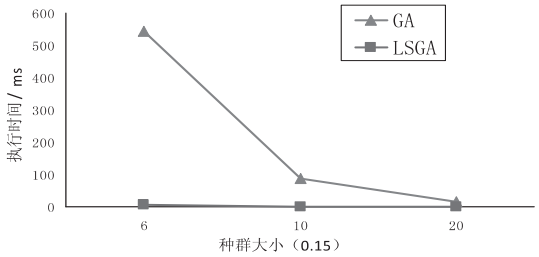

图 2　实验 2 运行时间对比图

4　结束语

文中对 LSGA 算法进行了理论分析,该算法能很好地保证种群的最大稳定性,提高搜索性能。首先从概率角度理论证明了 LSGA 算法优越于 GA 算法,然后将算法用于实例验证。结果表明,利用该算法能较快生成最小测试用例集,从而实现对测试目标的充分测试,提高测试效率,降低测试成本。

参考文献:

[1] Jones B F, Sthamer H H, Eyres D E. Automatic structural testing using genetic algorithms[J]. Software Engineering Journal, 1996, 11(5): 299-306.

[2] Hermadi I, Ahmed M A. Genetic algorithm based test data generator[C]//Proc of 2003 congress on evolutionary computation. [s. l.]: [s. n.], 2003: 85-91.

[3] Michael C C, McGraw G E, Schatz M A, et al. Genetic algorithms for dynamic test data generation[C]//Proc of 12th IEEE international conference on automated software engineering. [s. l.]: IEEE, 1997: 307-308.

[4] Jones B F, Eyres D E, Sthamer H H. A strategy for using genetic algorithms to automate branch and fault-based testing[J]. The Computer Journal, 1998, 41(2): 98-107.

[5] Wegener J, Baresel A, Sthamer H. Evolutionary test environment for automatic structural testing[J]. Information and Software Technology, 2001, 43(4): 841-854.

[6] Eugenia D, Javier T, Raquel B. Automated software testing using a metaheuristic technique based on tabu search[C]// Proc of 18th IEEE international conference on automated software engineering. [s. l.]: IEEE, 2003: 310-313.

[7] Johnson D S. Approximation algorithms for combinatorial problems[J]. Journal of Computer and System Sciences, 1974, 9(3): 256-278.

[8] Harrold M J, Gupta R, Soffa M L. A methodology for controlling the size of a test suite[J]. ACM Transactions on Software Engineering and Methodology, 1993, 2(3): 270-285.

[9] Chen T Y, Lau M F. A new heuristic for test suite reduction[J]. Information and Software Technology, 1998, 40(5-6): 347-354.

[10] Chen T Y, Lau M F. Heuristics towards the optimization of the size of a test suite[C]//Proceedings of the 3rd international conference on software quality management. Seville, Espagne: [s. n.], 1995: 415-424.

[11] Lee J G, Chung C G. An optimal representative set selection method[J]. Information and Software Technology, 2000, 42(1): 17-25.

[12] 聂长海,徐宝文. 一种最小测试用例集生成方法[J]. 计算机学报,2003,26(12):1690-1695.

[13] 马雪英,盛斌奎,叶澄清. 用遗传算法的测试用例最小化[J]. 计算机科学,2007,34(1):285-288.

[14] 全君林,陆　璐. 基于遗传算法测试用例集极小化研究[J]. 计算机工程与应用,2009,45(19):58-61.

[15] 申利民,高　洁. 基于遗传蚁群融合算法的测试用例最小化研究[J]. 计算机工程,2012,38(16):57-60.

[16] 刘　冬,靳蓓蓓,阙向红. 基于 LSGA 的最小测试用例集自动生成[J]. 微电子学与计算机,2011,28(12):115-118.

(上接第 85 页)

[15] Griffiths T L, Steyvers M. Finding scientific topics[J]. PNAS, 2004, 101(1): 5288-5235.

[16] Chang Chih-Chung, Lin Chih-Jen. LIBSVM: a library for support vector machines[EB/OL]. 2011. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[17] 杨　潇,马　军,杨同峰,等. 主题模型 LDA 的多文档自动文摘[J]. 智能系统学报,2010,5(2):169-176.

[18] Duda R O, Hart P E, Stork D G. Pattern classification[M]. 李宏东,姚天翔,译. 2nd ed. 北京:机械工业出版社,2003:508-576.

[19] 张明慧,王红玲,周国栋. 基于 LDA 主题特征的自动文摘方法[J]. 计算机应用与软件,2011,28(10):20-22.

[20] Lin J. Divergence measures based on Shannon entropy[J]. IEEE Transactions on Information Theory, 1991, 37(1): 145-151.

[21] Ruthven I, Lalmas M. A survey on the use of relevance feedback for information access systems[J]. Knowledge Engineering Review, 2003, 18(2): 95-145.

[22] 王　燕. 一种改进的 k-means 聚类算法[J]. 计算机应用与软件,2004,21(10):122-123.

[23] 王振振,何　明,杜永萍. 基于 LDA 主题模型的文本相似度计算[J]. 计算机科学,2013,40(12):229-232.

[24] 周昭涛. 文本聚类分析效果评价及文本表示研究[D]. 北京:中国科学院研究生院,2005.

[25] 姜　园,张朝阳,仇佩亮,等. 用于数据挖掘的聚类算法[J]. 电子与信息学报,2005,27(4):655-662.