基于 LDA 模型的文本聚类研究*

董婧灵[1,2]，李 芳[1,2]，何婷婷[1,2]，涂新辉[1,2]，万 剑[1,2]
[1]华中师范大学 计算机科学与技术系，湖北 武汉 430079
[2]国家语言资源监测与研究中心 网络媒体语言分中心，湖北 武汉 430079
E-mail: jinglingDong@google.com

摘 要：LDA（Latent Dirichlet Allocation）是近年来提出的一种具有文本主题表示能力的非监督学习模型。本文提出了一种基于 LDA 主题模型的文本聚类和聚簇描述方法。利用 LDA 模型挖掘隐藏在文本内的不同主题与词之间的关系，得到文本的主题分布；并将此分布作为特征融入到传统的向量空间模型来计算相似度进而对文本进行聚类；再利用主题信息对聚类结果进行聚簇描述。实验结果表明本文的方法能够明显地提高聚类的效果。

关键词：主题模型；LDA；文本聚类

Document Clustering Method Based on LDA Model

Dong Jing-ling[1,2], Li Fang[1,2], He Ting-ting[1,2], Tu Xin-hui[1,2], Wan Jian[1,2]
[1]Department of Computer Science, HuaZhong Normal University, Wuhan 430079
[2]Network Media Branch, National Language Resources Monitoring and Research Center, Wuhan 430079
E-mail: jinglingDong@google.com

Abstract: Latent Dirichlet Allocation (LDA) is an unsupervised model which exhibits superiority on latent topic modeling of text data in the research of recent years. This paper presents a method which improves effectiveness of text clustering by using LDA model. It can mine the hidden relationship between the different topics and the words from texts, and get the topic distribution. Then we mix the distribution as feature into the traditional VSM, and use the topics to describe the clustering result. Experimental results show that the method can improve clustering quality effectively.

Key words: topic model; Latent Dirichlet Allocation; text clustering

1 引言

随着 21 世纪科技的快速发展，人类周围的信息量迅猛膨胀。面对着海量杂乱无章的文本信息，如何提取出目标信息一直是自然语言处理领域研究的热点。聚类技术作为一种无监督的学习方法，可以将海量的未知文本信息划分成最合适的聚簇，使得在同一簇中的对象尽可能地相似，处于不同簇中的对象差异性尽可能地增大，进而从文本集合中发现信息的分布情况，缩小查询范围并且直接定位到目标信息。

聚类[1,2]技术作为一种无监督的学习方法，完全依靠数字的驱动是很难满足现实要求的，这对现有的聚类分析工具是一个很大的挑战。越来越多的学者意识到这点，并在不断地将已知知识加入到这个无监督的学习过程中[3]。有部分学者提出引入现有的语义知识库，比如 Hu 等[4]提出使用维基百科来创建概念库，将文本集映射到该概念空间模型上，从而结合词相似度、概念相似度和类别相似度来计算文本相似度。Anna[5]等指出可以通过维基百科来搜集构建文本的语料库，维基百科中的类别信息则用来辅助搜集内容关联性较强的条目，同时这种类别信息对文本架构也有一定的指导作用。

我们认为，除了引入外部语义知识外，还可以挖掘文本中自身蕴涵的潜在语义知识，这种语义知识直接来源于当前语料，可以更好地描述文档集合。现有的挖掘潜在语义知识的模型主要有

* 项目资助：国家自然科学基金重大研究计划课题（90920005）；国家自然科学基金项目（61003192）；973 国家重点基础研究发展计划课题（2007CB310804）；教育部哲学社会科学研究重大课题攻关项目（08JZD0032）；教育部/国家外国专家局高等学校学科创新引智计划课题（B07042）；湖北省自然科学基金计划项目（2009CDB145）；武汉市晨光计划项目（201050231067）；华中师范大学中央高校基本科研业务费项目（CCNU10A02009, CCNU10C01005）。

3 种：LSA[6]、PLSA[7]、LDA[8]。比如 Tomonari[9]等将 PLSA 与 LDA 等引入到日文文本聚类中，直接利用生成模型来取缔普通的向量空间模型，最后经过比较发现两者各有优缺点。从模型自身的角度看，LDA 有着突出的优点：首先 LDA 模型是完全概率生成模型，具有丰富的内在结构，并且可以利用成熟有效的概率算法来训练和使用模型；第二，LDA 模型参数空间的规模是 K×N（K 是隐含主题的数量，N 是词表中词的数量），与文本集规模无关，因此 LDA 更适合在大规模语料库上构造模型。

因此，本文提出了一种基于 LDA 主题模型的文本聚类和聚簇描述方法。首先，利用 LDA 模型对文本进行建模，挖掘隐藏在文本内的不同主题与词之间的关系，得到文本的主题分布；并将此分布作为特征融入到传统的向量空间模型来计算相似度，进而得到文档集合的相似度矩阵，在此基础上对文本进行聚类；最后，利用 LDA 建模过程中得到的主题分布信息对聚类结果进行聚簇描述。

2 LDA 主题模型

隐含狄利克雷分配[8]（LDA，Latent Dirichlet Allocation）是近年来这方面发展起来的一种重要的离散数据集合的建模方法。它基于一个常识性假设：文档集合中的所有文本均共享一定数量的隐含主题。基于该假设，它将整个文档集特征化为隐含主题的集合，而每篇文本被表示为这些隐含主题的特定比例的混合。LDA 是一个生成概率模型，它假设语料中的每一篇文本的生成过程如下：

Ⅰ. 选择 N～Poisson (ξ)；

Ⅱ. 选择 θ ～Dir (α)；

Ⅲ. 对每一个词 w_n ：

i. 选择一个 topic zn～Multinomial (θ)；

ii. 从概率分布 $p(w_n| z_n, \beta)$中选择一个词 w_n, p 为在 topic z_n 下的一个多项式概率分布。

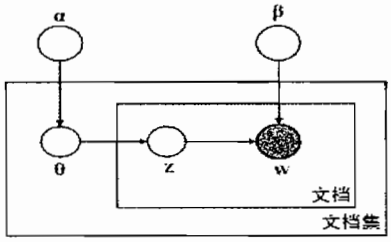


图 1　LDA 概率模型图

LDA 模型是由语料（文档集合）级参数 α 和 β 定义的。其中向量 α 反映了语料库中隐含主题间的相对强弱。矩阵 β 则刻画所有隐含主题自身在词语上的概率分布，其中元素 β_{ij} 表示第 i 个隐含主题生成第 j 个词的概率。 θ 是文本级的参数，表示每篇文本在主题上的分布。而 w 和 z 是词级的参数,每个词都要对应一次取值。

利用 LDA 模型对文档集合进行建模，可以得到两个矩阵：文本-主题分布矩阵和主题-词分布矩阵，从而可以挖掘出文本中潜在的语义知识。这种利用 LDA 模型提取隐含语义结构的方法已经成功地应用到很多相关领域[10,11]。

3 基于 LDA 模型的文本聚类

3.1 文本聚类简介

一般的文本聚类的过程可以分为三个阶段：预处理与文本建模阶段、聚类阶段、结果描述阶段，如图 2 所示。

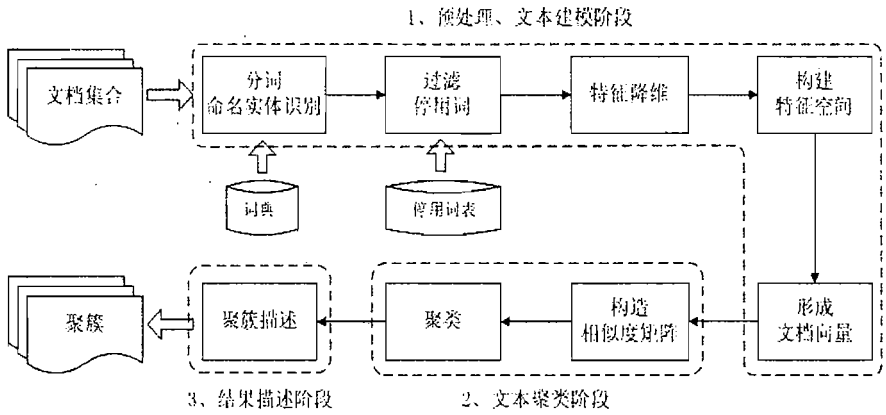


图2 文本聚类的一般体系结构

（1）预处理、文本建模阶段。预处理主要包括对文本进行分词、停用词过滤、命名实体识别等。文本建模是指对文本进行表示，如将文本表示为向量。

（2）文本聚类阶段。这一阶段主要是先计算两两文本间的相似度，得到相似度矩阵，再选择某个聚类算法在此基础上进行合并、划分等操作以实现聚类效果。

（3）结果描述阶段。聚类结果的描述更有助于结果的展示及用户的理解。目前常用的方法是将每个聚簇中所有文档的关联词语进行词频统计分析，进而找出最频繁出现的词语作为该聚簇的主题描述词。

以上可见，相似度计算是文本聚类中非常重要的一个步骤，对聚类结果的好坏有着直接的影响作用。但传统的相似度计算模型（如 VSM[12]）仅采取词频统计来表示文本，丢失了文本间大量的语义信息，从而影响了相似度计算的效果。因此，我们将采用 LDA 模型对文档集合进行建模，得到每个文本的主题分布向量，挖掘出潜在的语义知识，可以在一定程度上弥补单纯利用词频信息表示文本带来的信息丢失的不足。

3.2 基于 LDA 的文本相似度计算

LDA 主题模型是利用统计学的知识，分析文档集内部信息，将集合映射到基于隐含主题的特征空间上。根据该特征空间，我们提取了基于隐含主题的文本向量，结合加入 TF_IDF 权重的词向量，利用线性加权求和的方法，将两种文本表示向量进行有机融合，更有效计算地文本间的相似度。

对于每一篇文档 d_i，结合 TF_IDF 权重的词向量表示为 $d_{i\_VSM} = (w_1, w_2, w_3 \cdots w_n)$，其中 n 为 VSM 的维度。则文本 d_i, d_j 间基于词向量的余弦相似度为：

$$Sim_{VSM}(d_i, d_j) = \frac{d_{i\_VSM} * d_{j\_VSM}}{|d_{i\_VSM}| * |d_{j\_VSM}|}$$

基于 LDA 模型的主题向量表示为 $d_{i\_LDA} = (t_1, t_2, t_3 \cdots t_K)$，其中 K 为主题空间的维度。则文本 d_i, d_j 间基于隐含主题向量的余弦相似度为 $Sim_{LDA}(d_i d_j)$。

我们将两种文本相似度进行线性结合，计算公式如下：

$$Sim(d_i, d_j) = \lambda * Sim_{VSM}(d_i, d_j) + (1-\lambda) * Sim_{LDA}(d_i, d_j) \lambda \in (0,1)$$

其中 λ 为线性参数，表示 VSM 模型与 LDA 主题模型按一定比例的加权求和。

3.3 聚簇描述阶段

聚簇描述是结果展示的重要部分，其核心在于针对每个聚簇概括出定位类别的主题词，以便

更好的表示聚簇内容，帮助用户抓住核心思想。结合 LDA 生成的三层模型，我们采用如下步骤迭代进行。

Ⅰ. 针对聚类结果中的每篇文档 d_i，据文本-主题模型查找出占最大比重的隐含主题 $Topic_{max}(d_i)$；

Ⅱ. 统计每个聚簇中所有文档的 $Topic_{max}$，定位出每个聚簇中占最大比重的 $Topic_{key}$；

Ⅲ. 根据每个聚簇的 $Topic_{key}$，查找主题-词模型以及主题词列表，筛选出前三个主题描述词。

4 实验设计与结果分析

本文采用的聚类测试算法是传统的 K-means 方法[11]，为解决该算法中初始化的瓶颈问题，我们人工指定初始聚类中心，以便创造平衡公正的实验环境。

实验评估的指标采用 micro_F1 和 F1[13]。micro_F1 用来评价各聚类算法的综合性能。F1 则评价各聚类算法在各个类别上的聚类性能。

4.1 语料选择

我们分别在中英文语料上进行了测试。中文采用的是复旦中文语料库，为保证实验平衡性，实验抽取了其中 5 个子集，每个类 100 篇文本，类别分别是：C3-Art、C7-History、C19-Computer、C34-Economy、C39-Sports。词表大小 N 为 28096。英文则采用 Newsgroup 英文语料库，同样平均抽取 5 个子集的 500 篇文本，类别分别是 comp.os.ms-windows.misc.c、comp.sys.ibm.pc.hardware.d、rec.sport.baseball.j、sci.space.o、talk.politics.misc.s。词表大小 N 为 19126。

4.2 实验步骤

首先对文本进行预处理，将文本向量化，利用 VSM 模型计算相似度 $Sim_{VSM}(d_i,d_j)$；然后对文本进行 LDA 模型建模，得到文档的主题分布，计算相似度 $Sim_{LDA}(d_i,d_j)$；再将两者进行线性组合，得到文本相似度矩阵；最后，利用 K-means 算法进行文本聚类，并对聚类结果进行描述。

LDA 模型的建模过程如图 3 所示。建模过程中的参数估计利用 MCMC 方法中的 Gibbs 抽样算法[14]，具体设置 topic 的个数 K=50、$\alpha = 50/K$、$\beta = 0.01$，迭代次数均为 2000 次。其中主题数 K 的取值依次由 10 迭代到 200，经过多次实验，这里只给出效果最好的情况。

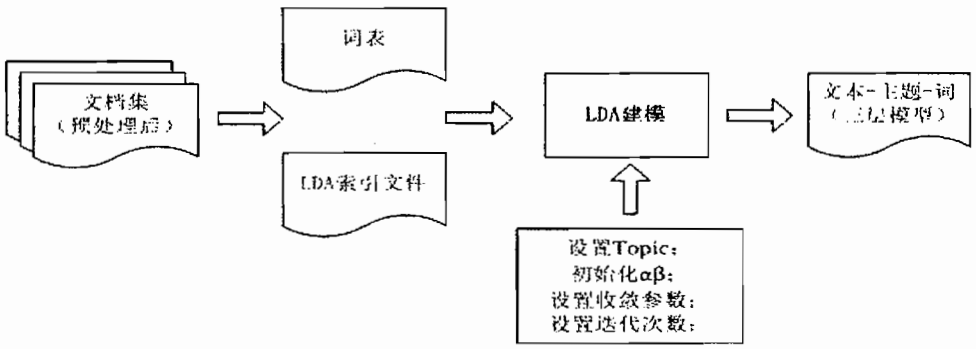
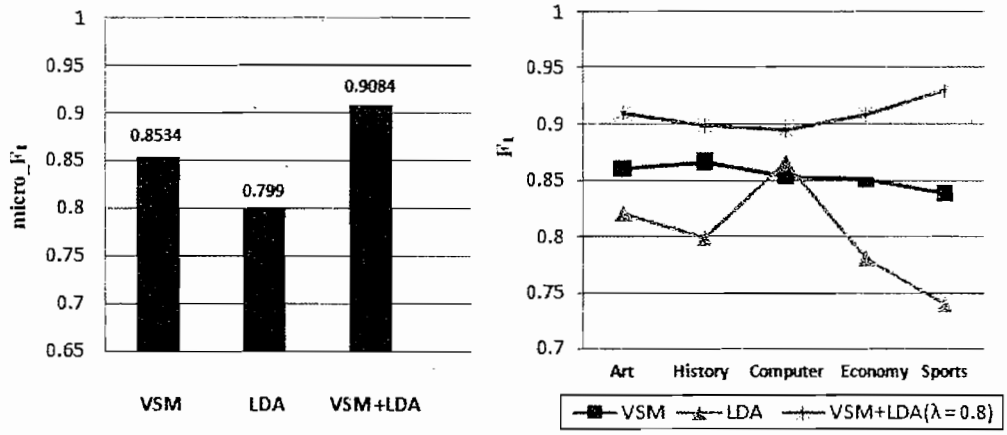


图 3 LDA 主题模型建模过程

4.3 实验结果及分析

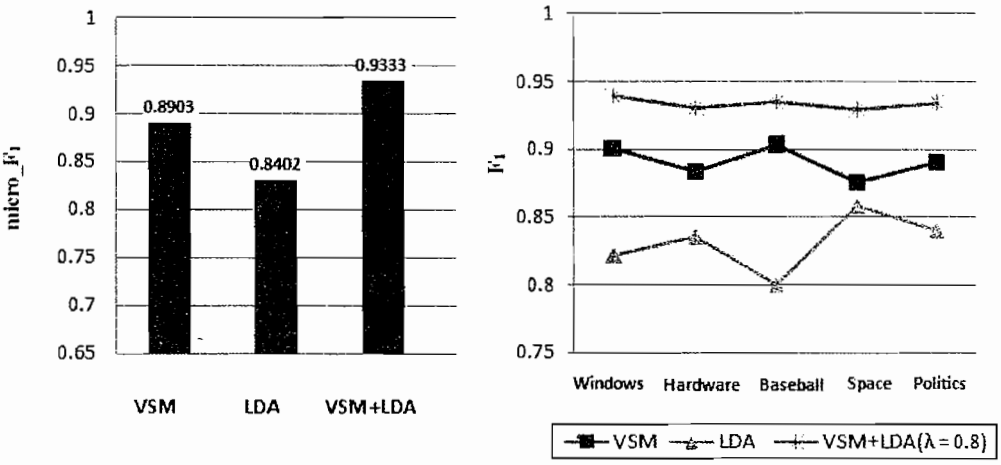
（1）聚类结果

中文和英文语料的聚类结果评估分别如图 4、图 5 所示。



中文语料基于微平均指标 micro_F1 的比较　　各类别上基于指标 F1 的比较

图 4



英文语料基于微平均指标 micro_F1 的比较　　各类别上基于指标 F1 的比较

图 5

图中的 VSM 指仅将 $Sim_{VSM}(d_i,d_j)$作为文本的相似度进行聚类；LDA 指将 $Sim_{LDA}(d_i,d_j)$看作文本的相似度进行聚类；VSM+LDA 是指利用本文的方法，将 VSM 和 LDA 融合起来计算文本的相似度进行聚类。实验过程中 λ 的值依次从 0.1 取到 0.9，多次试验发现，当 λ=0.8 时聚类质量最高。

可以看出，单独使用 LDA 模型的聚类效果很差，VSM 和 LDA 二者的恰当结合却可以明显地提高聚类的效果，在 VSM 模型的基础上分别提高了 5.5%和 4.3%（85.34%到 90.84%，89.03%到 93.33%），在 LDA 模型的基础上分别提高了 10.84%和 9.31%（79.99%到 90.84%，84.02%到 93.33%）。这是因为 LDA 模型只考虑了文本的主题分布，而主题向量的维度为 50，仅利用这样的低纬向量来计算文本相似度，必然丢失大量的信息，区分文本的力度是不够的。而 VSM 模型仅利用词频建立向量，同样也会丢失部分语义信息。但是将二者结合起来的 VSM+LDA 模型，则从主题和词语两个方面来衡量文本间的相似度，综合它们各自的优势，互相弥补不足，从而保证了聚类的效果。

（2）聚簇描述结果

聚类描述结果如表 1、表 2 所示，可以看出采用基于 LDA 模型的聚簇描述方法，能够准确地概括出结果主题，让聚簇结果更加直观。

表1 复旦中文聚类结果描述

类别	主题描述词
C5-Education	教育、改革、社会
C7-History	历史、发展、理论
C19-Computer	计算机、系统、网络
C34-Economy	经济、企业、市场
C39-Sports	体育、运动、比赛

表2 Newsgroup 英文聚类结果描述

类别	主题描述词
comp.os.ms-windows.misc.c	computer、windows、system
comp.sys.ibm.pc.hardware.d	hardware、machine、driver
rec.sport.baseball.j	sport、baseball、competition
sci.space.o	space、electronic、engineer
talk.politics.misc.s	politics、speech、conference

5 总结

本文将 LDA 主题模型引入到文本聚类领域，主要表现在文本建模、文本相似度计算以及聚簇描述三个方面。文本建模方面是利用了 LDA 模型的特性，在原本机械统计词频的基础上加入了文本的深层语义知识，从而让聚类过程更加精准，降低错误率。文本相似度计算方面则将常用的 VSM 模型与 LDA 主题模型进行一定比例的线性组合，建立多个文本特征空间，增强文本的向量表示，从而提高文本聚类的质量。聚簇描述则让聚类结果更加直观。在复旦中文语料库和 Newsgroups 英文语料库的实验表明，该方法能够明显地提高聚类的效果。

我们未来拟开展的研究工作包括：（1）如何进一步利用 LDA 主题模型，更好的表示文本特征，更深层的挖掘出文本信息；（2）LDA 模型是从文档集内部获取语义知识，如何利用外部语义知识库提高文本聚类质量。

参 考 文 献

[1] Salton G. Automatic Text Processing [M]. Boston: Addison Wesley Longman Publishing Company, 1988.

[2] Goldsmith M, Salami M. A Probabilistic Approach to Full Text Document Clustering [R].Technical Report ITAD 133 MS-98-044, SRI International, 1988.

[3] 景丽萍，恽佳丽，于剑. 领域知识在文本聚类过程中遇到的机遇与挑战[J]. 计算机工程与科学 2010.6: 88-91.

[4] Xiaohua Hu, Zhang X, et.al. Exploiting Wikipedia as External Knowledge for Document Clustering[C]. Process of the ACM SIGKDD 2009: 389-396.

[5] Anna Huang. Milne D, Frank E, et.al. Clustering Documents using a Wikipedia-Based Concept Representation[C]. Process of the PAKDD 2009: 628-636.

[6] Chen L, Tokuda N, Nagai A. A New Differential LSI Space-based Probabilistic Document Classifier. Information Processing Letters, 2003, 88(5): 203-212.

[7] Thomas Hofmann. Probabilistic Latent Semantic Indexing[A]. SIGIR 1999: 50-57.

[8] T D. Blei, A. Ng, M. Jordan. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, January 2003, 3:993-102.

[9] Tomonari Masada, Senya Kiyasu. Comparing LDA with pLSI as a dimensionality reduction method in document clustering[A]. LKR 2008, LNAI 4938, pp. 13-26, 2008.

[10] Wei X, Croft W B. LDA-based document models for adhered retrieval. Proceedings of the 29th SIGIR Conference. 2006: 178-185.

[11] Thorsten Brants , Ioann is Tsochantarid is Topic-based document segmentation with probab ilistic latent semantic analysis [A]. Proceedings of the eleventhinte rnational Conference on Information and knowledge management USA. 2002: 211-218.

[12] Han Jia Wei, Kamber M. Data Mining: Concepts and Techniques (2nd Edition)[M]. San Francisco: Morgan Kaufmann Publishers, 2006.

[13] Michael Steinbach, George Karypis, Vipin Kumar. A Comparison of Document Clustering Techniques [A]. Department of Computer Science and Engineering, University of Minnesota. Technical Report 00-034, 2000.

[14] Sean Borman, The Expectation Maximization Algorithm A short tutorial.