基于LDA主题模型的文本相似度计算

王振振 何 明 杜永萍
(北京工业大学计算机学院 北京 100124)

摘 要 LDA(Latent Dirichlet Allocation)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。提出了一种基于LDA主题模型的文本相似度计算方法,该方法利用LDA为语料库建模,利用MCMC中的Gibbs抽样进行推理,间接计算模型参数,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,并以此分布来计算文本之间的相似度,最后对文本相似度矩阵进行聚类实验来评估聚类效果。实验结果表明,该方法能够明显提高文本相似度计算的准确率和文本聚类效果。

关键词 主题模型,LDA,文本相似度,Gibbs抽样
中图法分类号 TP301 **文献标识码** A

Text Similarity Computing Based on Topic Model LDA

WANG Zhen-zhen HE Ming DU Yong-ping
(Department of Computer Science, Beijing University of Technology, Beijing 100124, China)

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 presented a method which improves text similarity calculation by using LDA model. This method models corpus and text with LDA. Parameters are estimated with Gibbs sampling of MCMC and the word probability is represented. It can mine the hidden relationship between the different topics and the words from texts, get the topic distribution, and compute the similarity between the text. Finally, the text similarity matrix clustering experiments are carrieel out to assess the effect of clustering. Experimental results show that the method can improve the text similarity accurate rate and clustering quality effectively.

Keywords Topic model, Latent Dirichlet Allocation(LDA), Text similarity, Gibbs sampling

1 引言

互联网作为一个分布式的、开放的信息平台近年来得到飞速发展,互联网上的信息也以指数级的方式增长,其中文本数据一直占据着重要地位。如何有效地从海量的文本数据中挖掘出有用的信息成为了当前的迫切需求。

文本相似度计算是各种文本挖掘技术的基石,有了文本相似度的定义就有了各种文本比较的理论依据。比如信息检索任务就可以看成是检索文本与被检索文本之间的一种相似度度量;而文本分类和文本聚类技术则运用了文本之间的离散度或者相似度概念;在文本自动摘要研究中,需要记住文本中句子之间的相似度问题;在自动问答系统中也需要计算问题与答案之间的相似度。由此可见,文本相似度的度量已经成为文本挖掘最核心的问题之一。

向量空间模型(VSM)[1]是信息检索领域最为经典的分析模型之一。其中基于TF-IDF的向量空间模型文本相似度计算方法是使用最广泛的文本相似度计算方法,这种方法以词在文本中出现频率以及在文本集中出现该词的频率来表征词的权重,通过计算向量之间的余弦值来计算文本的相似度。

文本是自然语言的载体,其自身必然包含着自然语言的复杂性,使用VSM必然不能完全建模自然语言的复杂性问题。文本的语义问题只是其中的一种,比如关于"苹果"和"手机"这两个词,其可能出现在两篇IT领域的文本中,但是由于词项不匹配,因此这两个词的相似度就很低;而当"苹果"一个出现在IT文章中,一个出现在水果文章中,以为这两个词一致,因此就把它们看成是相似的。此外用向量空间模型来表示文本数据,其数据空间也是极度高维且稀疏的,特别是对于中文文本数据来说更是如此,虽然常用汉字也就5000个左右,但是与由5000个汉字组成的词就会有十万百万个之多。

针对上述方法存在的缺陷,本文提出了将LDA主题模型[2]应用到文本相似度计算中。该方法利用LDA模型对文本集进行建模,即利用文本的统计特性,将文本语料库映射到各个主题空间,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,通过此分布来计算语料库的相似度矩阵。

到稿日期:2013-02-27 返修日期:2013-06-26 本文受国家自然科学基金(60803086),北京市自然科学基金(4123091),北京市教委科研计划(KM20110005013, KM200910005009)资助。

王振振(1988—),硕士生,主要研究方向为数据挖掘、信息检索;何 明(1975—),男,副教授,硕士生导师,主要研究方向为数据库理论与技术、数据挖掘、信息检索等;杜永萍(1977—),女,副教授,硕士生导师,主要研究方向为自然语言处理、信息检索等。

2 相关工作

LDA(Latent Dirichlet Allocation)模型是 Blei 提出的一种对离散数据集(如文档集)建模的概率主题模型[2]。作者从几何学出发详细解释了其与 unigram 模型、混合 unigram 模型、LSI 模型及 PLSI 模型的联系与区别,与这些相对简单的潜变量模型相比,LDA 模型有着突出的优点:首先 LDA 模型是全概率生成模型,具有清晰的层次结构,较 unigram 模型与混合 unigram 模型基于每篇文档只由一个主题生成的假设其更符合实际情况;其次 LDA 模型在主题层与词层都引入了 Dirichlet[3] 先验参数,解决了 LSI 模型与 PLSI 模型中主题参数个数随训练文档数目增加而线性增加,从而导致过度拟合的问题,因此更适合处理大规模语料库。该模型一经提出便被广泛地应用在文本分类、文本建模、图像处理及信息检索等领域。

Blei 等人利用 LDA 对文本进行建模,然后将建模后的文本使用支持向量机(SVM)进行分类,在降维幅度达到 99% 的情况下提高了文本分类的准确度。刘振鹿等人[4] 应用 LDA 模型进行文本的潜在语义分析,将语义分布划分成低频、中频、高频语义区,以低频语义区的语义进行 Web 游离文本检测,以中、高频语义区的语义作为文本特征进行文本聚类,采用文本类别与语义互作用机制对聚类结果进行修正,获得了很好的聚类效果。曹娟等[5] 研究了 LDA 模型的最优化问题,证明当主题之间的相似度最小时模型最优的理论。李文波等[6] 提出了一种附加类别标签的 LDA 模型(Labeled-LDA),通过在传统 LDA 模型中融入文本类别信息,提高了该模型的分类能力,其可以计算出隐含主题在各类别上的分配量,从而克服了传统 LDA 模型用于分类时强制分配隐含主题的缺陷,有效改进了文本分类的性能。石晶等[7] 利用 LDA 为语料库及文本建模,采取背景词汇聚类及主题词联想的方式将主题词扩充到待分析文本之外,尝试挖掘隐藏于字词表面之下的文本内涵,提高了文本分析的效果。Xing 等人[8] 将 LDA 模型和语言模型相结合,并使用基于聚类的方法提高了检索的召回率。

3 LDA 模型在文本相似度计算中的应用

3.1 LDA 模型基本思想

LDA 模型是一种对离散数据集(如文档集)建模的概率主题模型,是一种对文本数据的主题信息进行建模的方法,通过对文档进行一个简短的描述,保留本质的统计信息,有助于高效地处理大规模的文档集。它有 3 层生成式贝叶斯网络结构[9],基于这样一种前提假设:文档是由若干个隐含主题构成,而这些主题是由文本中若干个特定词汇构成,忽略文档中的句法结构和词语出现的先后顺序[10]。

LDA 模型具有清晰的层次结构,依次为文档集合层、主题层和特征词层,其结构图如图 1 所示。

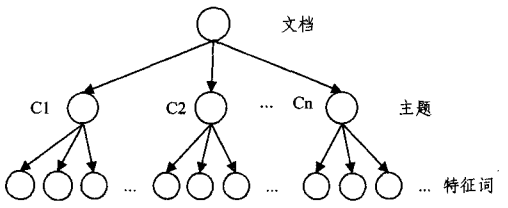

图 1 LDA 模型隐含主题的拓扑结构示意图

如图 2 所示,LDA 模型是典型的有向概率图模型[11],由参数(α,β)确定,α 反映了文档集合中隐含主题间的相对强弱,β 刻画所有隐含主题自身的概率分布。其中 θ_k 表示文档主题的概率分布,φ_k 表示特定主题下特征词的概率分布,M 表示文档集的文本数,K 表示文档集的主题数,N 表示每篇文档包含的特征词数。

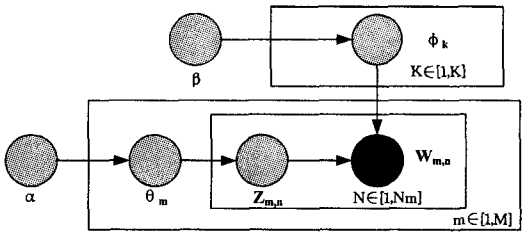

图 2 LDA 模型有向概率图

LDA 概率主题模型生成文本的过程[12] 如下:

(1)对于主题 z,根据 Dirichlet 分布 $Dir(\beta)$ 得到该主题上的一个单词多项式分布向量 φ;

(2)根据泊松分布 P 得到文本的单词数目 N;

(3)根据 Dirichlet 分布 $Dir(\alpha)$ 得到该文本的一个主题分布概率向量 θ;

(4)对于该文本 N 个单词中的每一个单词 Wn:

(4.1)从 θ 的多项式分布 $Multinomial(\theta)$ 随机选择一个主题 z;

(4.2)从主题 z 的多项式条件概率分布 $Multinomial(\varphi)$ 选择一个单词作为 Wn。

3.2 Gibbs 抽样

在构建 LDA 模型的过程中需要进行模型参数的估计,比较常用的估计方法主要有变分贝叶斯推理、期望传播算法和 Collapsed Gibbs 抽样等,基于 Gibbs 抽样的参数推理方法容易理解且实现简单,能够非常有效地从大规模文本集中抽取主题,因此,Gibbs 抽样算法成为当前最流行的 LDA 模型抽取算法。

在 LDA 模型中,最重要的两组参数分别是各主题下的词项概率分布和各文本的主题概率分布。本文参数估计利用 MCMC[13] 方法中的 Gibbs 抽样算法[14],可以看成是文本生成过程的逆过程,即在已知文本集(即生成的结果)的情况下,通过参数估计得到参数值。根据图模型,可以得到一篇文本的概率值为:

$$p(w|\alpha,\beta)=\int p(\theta\mid\alpha)\left(\prod_{n=1}^{N}\sum_{z_n}p(z_n|\theta)p(w_n|z_n,\beta)\right)\mathrm{d}\theta \tag{1}$$

所谓"Collapsed"是指通过积分避开了实际待估计的参数,转而对每个单词的主题进行采样,一旦每个单词的主题确定下来,参数就可以在统计频次后计算出来。因此,参数估计问题变为计算单词序列下主题序列的条件概率,其公式如下:

$$p(z_i=k|\vec{z}_{\neg i},\vec{w})=\frac{p(\vec{w},\vec{z})}{p(\vec{w},\vec{z}_{\neg i})}\propto\frac{n_{k,\neg i}^{t}+\beta_t}{\sum_{t=1}^{V}n_{k,\neg i}^{t}+\beta_t}(n_{m,\neg i}^{t}+\alpha_k) \tag{2}$$

式中,z_i 表示第 i 个单词对应的主题变量;$\neg i$ 表示不包括其中的第 i 项;n_k^t 表示 k 主题中出现词项 t 的次数;β_t 是词项 t 的 Dirichlet 先验;n_m^k 表示文本 m 出现主题 k 的次数;α_k 是主题 k 的 Dirichlet 先验。

一旦获得每个单词的主题标号,需要的参数计算公式可

由下面公式获得：

$$\varphi_{k,t}=\frac{n_k^t+\beta_t}{\sum_{t=1}^{V} n_k^t+\beta_t} \tag{3}$$

$$\theta_{m,k}=\frac{n_m^k+\alpha_k}{\sum_{k=1}^{K} n_m^k+\alpha_k} \tag{4}$$

式中，$\varphi_{k,t}$表示主题k中词项t的概率；$\theta_{m,k}$表示文本m中主题k的概率。

3.3 相似度计算

由于文本的主题分布是文本向量空间的简单映射，因此在文本的主题表示情况下，计算两个文本的相似度可以通过计算与之对应的主题概率分布来实现。由于主题是词向量的混合分布，因此有人使用KL (Kullback-Leibler)距离[15]作为相似度度量标准，KL距离如下所示：

$$D_{KL}(p,q)=\sum_{j=1}^{T} p_j \ln \frac{p_j}{q_j} \tag{5}$$

当对于所有的j，当$p_j=q_j$时，$D_{KL}(p,q)=0$。但是KL距离并不是对称的，即$D_{KL}(p,q)\neq D_{KL}(q,p)$，因此常常使用其对称版本：

$$D_\lambda(p,q)=\lambda D_{KL}(p,\lambda p+(1-\lambda)q)+(1-\lambda)D_{KL}(q,\lambda p+(1-\lambda)q) \tag{6}$$

当$\lambda=1/2$时，上述公式转变为JS距离[16]，JS距离的区间为[0,1]，本文以JS距离公式为标准来度量文本之间的相似度。

$$D_{js}(p,q)=\frac{1}{2}[D_{KL}(p,\frac{p+q}{2})+D_{KL}(q,\frac{p+q}{2})] \tag{7}$$

4 实验设计与结果分析

本文采用的聚类算法为传统的K-means算法[17]，实验评估的指标采用F度量值[18]来衡量文本的相似度。F度量值是信息检索中的一种组合查准率和召回率指标的平衡指标。

准确率$P(i,j)$和召回率$R(i,j)$可分别定义为：

$$P(i,j)=\frac{n_{ij}}{n_j} \tag{8}$$

$$R(i,j)=\frac{n_{ij}}{n_i} \tag{9}$$

式中，n_i是类别i的文本数目，n_j是聚类j的文本数目，n_{ij}是聚类j中隶属于i的文本数目。

对应的F度量值定义为：

$$F(i,j)=\frac{2*P(i,j)*R(i,j)}{P(i,j)+R(i,j)} \tag{10}$$

全局聚类的F度量值定义为：

$$F=\sum_i \frac{n_i}{n}\max_j(F(i,j)) \tag{11}$$

4.1 语料选择

本文只在中文语料上进行了测试，采用的是复旦中文语料库，为保证实验的平衡性，实验抽取了其中8个子集，分别是：C3-Art、C11-Space、C19-Computer、C31-Environment、C32-Agriculture、C34-Economy、C38-Politics、C39-Sports。每个类分别包含400篇文本，共3200篇文档。

4.2 实验步骤

首先对文本进行预处理，包括分词、去停用词等，将文本向量化，表示为文档-特征词矩阵；然后对上述文档矩阵构建LDA模型，得到文档的主题概率分布，根据JS距离计算文档之间的相似度，得到相似度矩阵；最后利用K-means算法对文本进行聚类，分析聚类结果来评价文档相似度计算的准确性。

LDA建模过程中，参数估计利用MCMC方法中的Gibbs抽样算法，具体设置topic的初始个数$K=50$、$\alpha=50/K$、$\beta=0.01$，Gibbs抽样的迭代次数为1000次。其中主题数K的取值依次为50，100，直到400，利用不同主题数进行多次聚类实验，获得最优主题数K。

4.3 实验结果分析

从图3中可以看出当主题数为250时F度量值最高，因此我们选择主题数为250。选取最优主题数后就可以得到利用LDA模型计算文本相似度的最终聚类结果，本文还通过K-means算法对比了基于TF-IDF的向量空间模型的相似度计算方法，其对比图如图4所示。

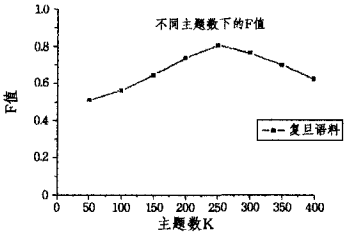

图3 不同主题数的聚类结果

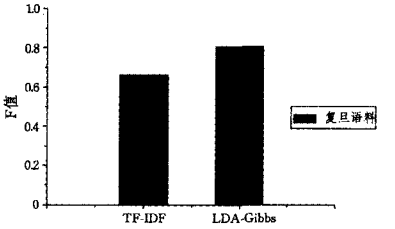

图4 不同方法聚类结果比较

从图4中可以看出，本文方法比基于TF-IDF的相似度计算方法的聚类效果更好。

结束语 本文将LDA主题模型应用到文本相似度计算中，主要表现在文本建模、文本相似度计算方面。文本建模利用了LDA模型的特性，加入了文本的深层语义知识，从而使聚类过程更加精准。在文本相似度计算方面，利用LDA建立了文本主题空间，增强了文本的向量表示，大大缩小了文档的维度，加快了计算速度，从而提高了聚类效果。在复旦中文语料库的实验表明，该方法可以明显提高文本相似度计算的准确率和文本聚类效果。

由于LDA非常容易扩展，本文后续的研究将在LDA模型的基础上继续探讨文本建模方法以及基于其上的文本挖掘，如文本分类、相似项挖掘等，这种模型对于数据挖掘、自然语言处理等学科具有重要意义。

参 考 文 献

[1] Salton G, Wong A, Yang C S. A Vector Space Model for Automatic Indexing[J]. Communications of the ACM, 1975, 18: 613-620

[2] Blei D, Ng A, Jordan M. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3: 993

[3] 徐谦，周俊生，陈家骏. Dirichlet过程及其在自然语言处理中的

应用[J]. 中文信息学报,2009(5):125
[4] 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):60-67
[5] 曹娟,张勇东. 一种基于密度的自适应最优 LDA 模型选择方法[J]. 计算机学报,2008,31(10):1780-1788
[6] 李文波,孙乐,黄瑞红,等. 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,范猛,李万龙. 基于 LDA 模型的主题分析[J]. 自动化报,2009,36:1586-1593
[8] Wei Xing, Croft W B. LDA-Based Document Models for Ad-hoc Retrieval[C]// SIGIR'06. Seattle, WA, USA, August 2006
[9] Friedman N, Geiger D, Goldszmidt M. Bayesian Network Classifiers[J]. Machine Learning, 1997, 2:131
[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,2011,13:29-38
[11] 徐戈,黄厚峰. 自然语言处理中主题模型的发展[J]. 计算机学报,2011,34(8):1423-1437
[12] 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,2011,10:215
[13] Doucet A, Godsill S, Andrieu C. On sequential Monte Carlo sampling methods for Bayesian filtering[J]. Statistics and Computing, 2000, 3:197
[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,2011,2:89-118
[15] Duda R O, Hart P E, Stork D G. Pattern Classification(2ed)[M]. 李宏东,姚天翔,等译. 机械工业出版社,2003:508
[16] Lin J. Divergence measures based on Shannon entropy[J]. IEEE Transactions on Information Theory, 1991, 37(14):145
[17] 王燕. 一种改进的 k-means 聚类算法[J]. 计算机应用与软件,2004,10(3):122
[18] 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

(上接第 204 页)

需证书管理;同时 LT-CA 私钥采用(t,n)门限机制由 n 个认证服务器共享,各服务器对持有的子密钥进行周期性刷新,使更新后的各子密钥仍然共享同一个秘密,系统的入侵容忍性极大提高,可抵御无线环境下易于实施的多种攻击;原型实现及仿真实验表明,门限机制并没有明显增加系统的计算量及负载,时间代价在可接受范围内,系统安全性显著增强,适用于资源受限的 WMN 网络。

参 考 文 献

[1] Qi Ji, Zhao Yi, Wang Xing-ming, et al. Security authentication and an undeniable billing protocol for WMNs[C]// Sterritt R. Proceedings of International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery. Huang Shan, China: IEEE Publisher, 2010:266-269
[2] Durahim A O, Savas E. A2-MAKE: An efficient anonymous and accountable mutual authentication and key agreement protocol for WMNs[J]. Ad Hoc Networks, 2011, 9(5):1202-1220
[3] Boudguiga A, Lauren URENT M. Key-escrow resistant ID-based authentication scheme for IEEE 802.11s Mesh Networks [C]// Kingston D. Proceedings of IEEE Wireless Communications and Networking Conference (WCNC). Quintana Roo, Mexico: IEEE Publisher, 2011:784-789
[4] Shamir A. How to share a secret [J]. Communications of the ACM, 1979, 22(11):612-613
[5] Blakley G R. Safeguarding cryptographic keys[C]// Smith M. Proceedings of the National Computer Conference. New York, USA: IEEE Publisher, 1979:313-317
[6] Kim J, Bahk S. Design of certification authority using secret redistribution and multicast routing in wireless mesh networks [J]. Computer Networks, 2009, 53(1):98-109
[7] Yang Kan, Jia Xiao-hua, Zhang Bo, et al. Threshold key redistribution for dynamic change of authentication group in Wireless Mesh Networks[C]// LIANG J. Proceedings of IEEE Global Telecommunications. Miami, USA: IEEE Publisher, 2010:1156-1151
[8] Chai Zhen-chuan, Cao Zhen-fu, Lu Rong-xing. Threshold password authentication against guessing attacks in Ad hoc networks[J]. Ad-hoc Networks, 2007, 5(7):1046-1054
[9] Dong Xiao-lei, Wang Li-cheng, Cao Zhen-fu. New public key cryptosystems with lite certification authority[EB/OL]. http://ePrint.iacr.org/2006/154, 2013-3-16
[10] 潘耘,王励成,曹珍富,等. 基于轻量级 CA 的无线传感器网络密钥分配方案[J]. 通信学报,2009,30(3):130-134
[11] Dong Xiao-lei, Wei Li-fei, Zhu Hao-jin, et al. EP^{2}DF: an efficient privacy-preserving date-forwarding scheme for service-oriented vehicular Ad Hoc networks[J]. IEEE Transactions on Vehicular Technology, 2011, 60(2):580-591
[12] Nenal K. Elliptic curve cryptosystems[J]. Mathematics of Computation, 1987, 48(13):203-209
[13] Roman R, Alcaraz C. Applicability of public key infrastructures in Wireless Sensor Networks[C]// LOPEZ J. Proceedings of European PKI Workshop: Theory and Practice. Palma de Mallorca, Spain: Springer LNSC4582, 2007:313-320
[14] He B, Agrawal D P. An identity-based authentication and key establishment scheme for multi-operator maintained Wireless Mesh Networks[C]// Nayak A, Stojmenovic I. Proceedings of Mobile Ad Hoc and Sersor Systems. San Francisco, USA: IEEE Publisher, 2010:71-78
[15] Lin Xiao-dong, Lu Rong-xing, Ho Pin-han, et al. TUA: a novel compromise-resilient authentication architecture for Wireless Mesh Networks[J]. IEEE Transactions on Wireless Communications, 2008, 7(4):1389-1399
[16] Eissa T, Razak S A, Ngadi M D. Towards providing a new light-weight authentication and encryption scheme for MANET[J]. Wireless Network, 2011(17):833-842
[17] Barr R. Swans-scalable wireless Ad hoc network simulator user's guide[EB/OL]. http://www.isi.edu/nsnam/ns, 2013-03-21
[18] Barreto P S L M, Kim H Y, Lynn B, et al. Efficient algorithms for pairing-based cryptosystems[C]// Yung M. Proceedings of the 22nd annual international cryptology conference on advances in cryptology. Santa Barbara, USA: Springer, 2002:354-368
[19] Gura N, Patel A, Wander A, et al. Comparing elliptic curve cryptography and RSA on 8bit CPUs[C]// Joye M, Quisquater J J. Proceedings of Workshop on Cryptographic Hardware and Embedded Systems. Boston, USA: Springer, 2004:119-132