基于 LDA 的中文文本相似度计算

孙昌年[1,2],郑 诚[1,2],夏青松[1,2]
(1. 安徽大学 计算机科学与技术学院,安徽 合肥 230039;
2. 教育部计算智能与信号处理重点实验室,安徽 合肥 230039)

摘 要:传统基于 TF-IDF 的向量空间模型的文本相似度计算存在高维、数据稀疏、缺乏语义和维度未归一等问题,基于其上的语义扩展的 TF-IDF 向量空间模型中部分解决了语义问题,但是其基于词典的词语相似度计算限制了其应用范围。提出了一种基于潜在狄利克雷分配(Latent Dirichlet Allocation, LDA)的文本相似度计算方法,LDA 模型可以在没有词典的情况下解决上述所有问题,通过吉比斯抽样方法将文本建模到主题空间,然后使用 JS(Jensen-Shannon)距离来计算文本相似度。通过聚类实验表明该方法取得了较高的 F 值。

关键词:向量空间模型;文本相似度;自然语言处理;潜在狄里克雷分配;主题模型
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2013)01-0217-04
doi:10.3969/j.issn.1673-629X.2013.01.053

Chinese Text Similarity Computing Based on LDA

SUN Chang-nian[1,2],ZHENG Cheng[1,2],XIA Qing-song[1,2]
(1. School of Computer Science and Technology, Anhui University, Hefei 230039,China;
2. Key Lab. of Intelligent Computing & Signal Processing, Mini. of Edu., Anhui Univ., Hefei 230039,China)

Abstract:Text similarity calculation based on traditional TF-IDF vector space model exists high dimensional sparse data,lack of semantic and dimension normalization,the TF-IDF vector space model based on its semantic extension is to solve the partial problem of semantic, but its word similarity computation based on dictionary limits its application scope. Proposed a text similarity computing method based on potential Dirichlet distribution (Latent Dirichlet Allocation,LDA),LDA model can solve all these problems in no dictionary,through the Gibbs sampling method,the text modeling to subject space,and then use JS (Jensen-Shannon) distance computing text similarity. The clustering experiment results show that this method can achieve high F value.

Key words:vector space model;text similarity;natural language processing;latent Dirichlet allocation;topic model

0 引 言

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的频率以及在文本集中出现的该词的频率来表征词的权重,通过计算向量之间的余弦相似度来计算文本的相似度。忽略了文本中词项的含义,因而也就无法分辨出同义词与多义词,而同义词与多义词对于计算文档相似度具有重要的意义。此外,对于大多数文本数据集而言,词项的数目和文本数目通常都很大,而采用词频向量模型,必须将文本表示为词项数目与文本数目大致相当的矩阵,矩阵中的行数位文本集中的词项数目,大小一般在几万维,而矩阵中的列数目则为文本集中的文本数量,因而有着非常高的维度并且是极度稀疏的,最终导致了非常低效的计算。基于词项语义来考察文本相似度的方法在文本表示模型上多数沿用了词频向量模型,利用外部词典(如 WordNet、HowNet、同义词词林等)的引入来计算词项之间的相似度度量[1~4],这种引入外部词典的方法需要设计复杂的数据结构来提高计算效率,增加了系统设计的复杂度,而且又无法解决词典中未登录词的语义问题,而且这种方法很难移植到没有语义词典的应用中,方法的鲁棒性较差,再者没有针对文本表示的高维模型进行降维处理,也缺乏衡量文本间相似度的定义,导致基于词项语义相似度的文本计算方法局限性较大、难以扩展。

针对上述方法存在的缺陷,文中使用主题模型中的 LDA 模型对文本进行建模,该方法利用文本的统计特性,能有效降低文本表示维度,同时又能解决同义词

收稿日期:2012-04-16;**修回日期**:2012-07-21
基金项目:安徽省自然科学基金(06060716);安徽大学研究生学术创新研究(YQH090047)
作者简介:孙昌年(1985-),男,硕士研究生,研究领域为文本数据挖掘;郑 诚,副教授,研究领域为数据挖掘、语义网。

和多义词问题,并且无需引用外部词典的相似度计算方法,这种方法绕开了外部词典的引入,因而避免了词典中未登录词无法得到语义的问题。

1 相关工作

自然语言处理中的主题模型起源于隐性语义索引(Latent Semantic Indexing,LSI)[5]。LSI 并不是概率模型,因此也算不上是主题模型,但基于其主要思想,Hofmann 提出了概率隐性语义索引(Probabilistic Latent Semantic Indexing,pLSI)[6],pLSI 模型被看做是一个真正意义上的主题模型。此后 Blei 等人提出隐性狄里克雷分配(Latent Dirichlet Allocation,LDA)[7]进一步完善了主题模型。文中计算文本相似度的方法主要是基于 LDA 模型。

LDA 模型通过类似于词聚类的办法将相似词聚类为一个个主题,使得主题以及主题之间具有语义上的意思,对于在同一个主题中的词项一般具有近义词特性,而在不同主题中的同一个词则具有多义词特性,从而在文本相似度计算机的过程中免去了计算词项之间的相似度,而利用文本的主题分布可以计算文本之间的相似度,而且其计算在不需要外部词典的情况下,其计算结果也具有语义效果,比起基于外部词典的方法算法更具鲁棒性。基于 LDA 的主题模型是一种生成模型,图 1 是使用板图的方法对 LDA 模型的表示。

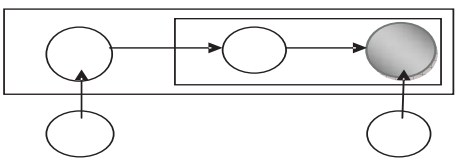

图 1 LDA 的模型图表示

图中方框表示循环,右上角的 M、N 表示循环次数,阴影区域代表观测变量,文中表示文本中的单词,空心节点表示隐含变量,箭头表示依赖关系。α 是 θ 的超参数,β 是 $K \times V$ 的参数集合,K 是主题个数,V 是词项个数,θ 表示某文本的主题概率分布,共 M 个,M 为文本个数,w 为单词,z 为 w 的主题标号。

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档的主题概率分布。参数估计可以看成是生成过程的逆过程: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:

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

在实际的应用中需要对 LDA 模型进行参数估计,其参数估计方法主要有变分贝叶斯推理、期望传播[8]和 Collapsed Gibbs Sampling[9]等,Gibbs 方法首先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[10],其公式如下:

$$p(z_i = k \mid \vec{z}_{\neg i}, \vec{w}) = \frac{p(\vec{w},\vec{z})}{p(\vec{w},\vec{z}_{\neg i})} \propto \frac{n^t_{k,\neg i} + \beta_t}{\sum_{t=1}^{V} n^t_{k,\neg i} + \beta_t}\left(n^k_{m,\neg i} + \alpha_k\right)$$

其中,z_i 表示第 i 个单词对应的主题变量;$\neg i$ 表示剔除其中的第 i 项;n^t_k 表示主题 k 出现词项 t 的次数;β_t 是词项 t 的 Dirichlet 先验;n^k_m 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式获得:

$$\varphi_{k,t} = \frac{n^t_k + \beta_t}{\sum_{t=1}^{V} n^t_k + \beta_t}$$

$$\vartheta_{m,k} = \frac{n^k_m + \alpha_k}{\sum_{t=1}^{K} n^k_m + \alpha_k}$$

其中 $\varphi_{k,t}$ 表示主题 k 中词项 t 的概率;$\vartheta_{m,k}$ 表示文本 m 中主题 k 的概率。因此,主要知道了每个单词的主题标号,那么就可以通过简单计数的方式对参数进行估计。

2 基于 LDA 的文本相似度

2.1 文本—词向量转换为文本—主题向量

基于 Gibbs 抽样的参数推理方法较容易实现,对其中的不必要的计算省去,得到下面的计算主题空间的算法,该算法参考了文献[11]中的算法,不同之处在于省去了对主题—词空间的计算过程。

算法描述如下:

```
Phi_Gibbs( {w⃗} , α , β , k )
{
n_m^(k) = n_m = n_k^(t) = n_k = 0 ;
for ( d = 1;d < m;d ++)
{
for( w = 1;w < n_m;w ++)
{
抽样主题标记 z_{m,n} = k ~ Mult(1/K) ;
n_m^(k) ++; n_m ++ ; n_k^(t) ++ ; n_k ++ ;
}
}
While( 1 )
{
for ( d = 1;d < m;d ++)
{
for( w = 1;w < Nm;w ++)
```

}

$n_m^{(k)}$ -- ; n_m -- ; $n_k^{(t)}$ -- ; n_k -- ;

抽样主题标记

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

$n_m^{(k)}$ ++ ; n_m ++ ; $n_k^{(t)}$ ++ ; n_k ++ ;

}

}

If (收敛或者达到迭代次数)

{

计算主题分布参数集 $\vartheta_{m,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^{K} n_m^{(k)} + \alpha_k}$

}

}

}

经过 Gibbs 抽样算法之后,就可以将文本的词向量空间映射为文本的主题向量空间。接下来就可以将该结果作为文本相似度计算的输入了。

2.2　计算文本相似度

由于文本的主题分布是文本向量空间的单纯形映射,所以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,因而使用 KL(Kullback - Leibler)[11]距离作为相似度度量标准,KL 距离如下所示:

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

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

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

容易证明 $D_\lambda(p,q) = D_\lambda(q,p)$ 。当 $\lambda = \frac{1}{2}$ 时,上述公式转变为 JS(Jensen-Shannon)距离:

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

文献[2]指出 JS 距离的区间为[0,1],文中以 JS 距离公式为标准来度量文本之间的相似度。

3　实　验

3.1　实验数据及度量标准

文中的实验数据是两个常用的文本分类数据集:复旦文本分类数据集、谭松波中文文本分类数据集 TanCorpV1.0。两个数据集特别是 TanCorp 的数据分布是严重偏斜的,由于文中不准备讨论数据偏斜的问题,所以为了有效验证文中的方法,采用以最小的类别文档数为标准裁剪数据集。复旦文本分类数据集裁剪后每个类别包含 200 篇文章,TanCrop 数据集裁剪后每个类别包含 150 篇文章。

实验采用 F-度量值来衡量文中提出的文本相似度。设 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,则查准率 $P(i,j)$ 和查全率 $R(i,j)$ 可分别定义为:

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

F 度量值定义为:

$$F = \sum_i \frac{n_i}{n} \max_j \left(\frac{2 \times P(i,j) \times R(i,j)}{P(i,j) + R(i,j)}\right)$$

其中 n 是文本集合中总的文本数目。通常 F 度量值越大,聚类效果越好。

3.2　实验结果

在对数据简单处理后,使用 FudanNLP 平台提供的分词技术对文本进行分词以及去除停用词。预处理之后的文本用词向量代替。预处理之后,将文本集做 LDA 处理,LDA 处理中先验超参数 α 和 β 的经验取值为 $\alpha = 50/K$, $\beta = 0.01$。K 的取值对于不同文本集而言取值不是固定的。通过实验的方式来确定主题数 K 的取值。确定的标准为聚类结果 F 值最高的主题数目,其中的聚类算法为 K-Means++算法,使用开源工具 lingPipe 中的 K-Means++实现。图 2 显示了不同主题数量对于聚类 F 值影响。

从图 2 中可以看出在主题为 200 时 F 值最高,因此在后续的比较中选择主题为 200。通过 K-Means++算法对比了基于 TF-IDF 的向量空间模型以及在 TF-IDF 基础上加入基于词典语义的向量空间模型的方法,实验对比结果如图 3 所示。

从图 3 中可以看出文中的方法比基本的 TF-IDF 相似度方法以及基于词典计算语义相似度的 IF-IDF 扩展方法的聚类效果要好。

4　结束语

文中提出了基于 LDA 主题模型来建模文档并使用 JS 距离来计算文档相似度的方法。由于在文本建模以及计算相似度的过程中没有使用任何外部词典,相似度计算的效果完全是根据不同文本集自身的特性,因而文中的计算方法其鲁棒性相对较好。另外,由于在使用 LDA 建模的过程中,对于文本向量的维度压缩比相当大,这给后来的相似度计算和聚类计算带来了相对大的计算效率的提升。实验表明了文中方法的有效性。

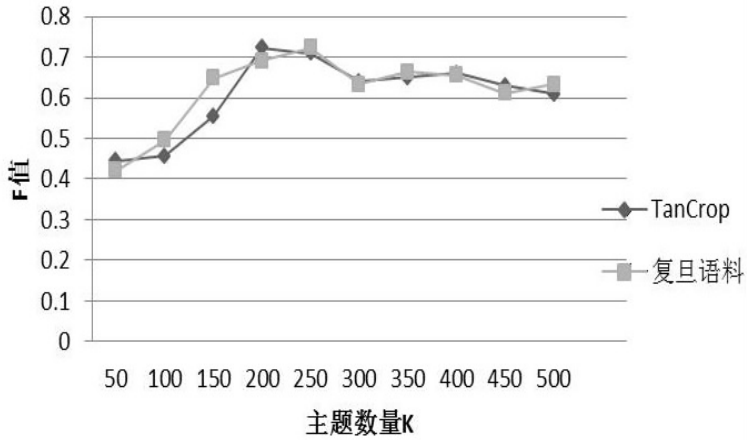

图 2 不同主题数量对聚类效果的影响

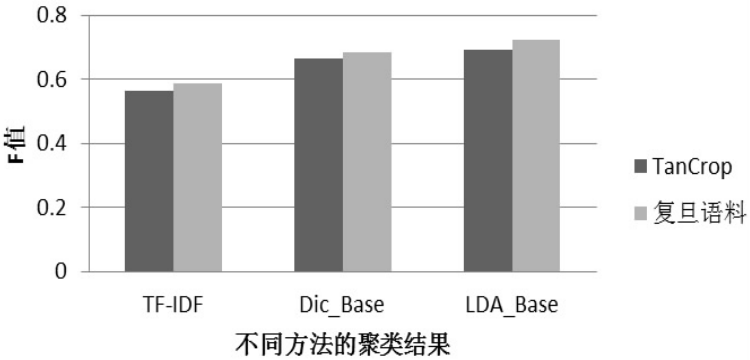

图 3 三种方法的效果比较

参考文献:

[1] Hotho A, Staab S, Stumme G. Wordnet improves text document clustering[C]//Proceeding of the Semantic Web Workshop at SIGIR-2003, 26th Annual International ACM SIGIR Conference. Toronto, Canada: [s. n.], 2003: 541-550.

[2] 李 峰,李 芳. 中文词语语义相似度计算-基于《知网》2000[J]. 中文信息学报,2007,21(3):99-105.

[3] 汪 敏,肖诗斌,王弘蔚,等. 一种改进的基于《知网》的词语相似度计算[J]. 中文信息学报,2008,22(5):84-90.

[4] 黄承慧,印 鉴,侯 昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法[J]. 计算机学报,2011(5):857-866.

[5] Deerwester S, Dumais T, Landauer G, et al. Indexing by latent semantic analysis[J]. Journal of the American Society of Information Science, 1990, 41(6): 391-407.

[6] Hofmann T. Probabilistic latent semantic indexing[C]//Proceedings of the Twenty-second Annual International SIGIR Conference. [s. l.]: [s. n.], 1999.

[7] Blei D, Ng A, Jordan M. Latent Dirichlet Allocation[M]//Advances in Neural Information Processing Systems 14. Cambridge, MA: MIT Press, 2002.

[8] Minka T, Lafferty J. Expectation propagation for the generative aspect model[C]//Proceeding of UAI 2002. Edmonton, Alberta, Canada: [s. n.], 2002: 352-359.

[9] Heinrich G. Parameter estimation for text analysis[R]. Darmstadt, Germany: [s. n.], 2009.

[10] Steyvers M. Probabilistic Topic Models[M]//Latent Semantic Analysis: A Road to Meaning. Mahwah: Lawrence Erlbaum Associates, 2007: 424-440.

[11] 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.

[12] Lin J. Divergence measures based on Shannon entropy[J]. IEEE Transactions on Information Theory, 1991, 37(14): 145-151.

(上接第 216 页)

Navigation and Control Conf. . Portland OR: [s. n.], 1999.

[3] Prandini M, Lygeros J, Nilim A, et al. Randomized Algorithms for Probabilistic Aircraft Conflict Detection[C]//Conference on Decision and Control-CDC. [s. l.]: [s. n.], 1999.

[4] Erzberger H, Paielli R A. Concept for Next Generation Air Traffic Control System[J]. Air Traffic Control Quarterly, 2002, 10(4): 355-378.

[5] Paielli R A, Erzberger H. Conflict probability estimation for free flight[J]. Journal of Guidance, Control and Dynamics, 1997, 20(3): 588-596.

[6] 罗世谦,冯子亮. 一种高效的中期冲突探测随机化算法[J]. 计算机应用与软件,2010,27(3):56-57.

[7] 李俊菊,宋万忠,梁海军,等. 中期冲突探测算法的研究与设计[J]. 计算机工程与设计,2010,31(20):4493-4494.

[8] 吴舜歆,彭 炜,李瑞芳. 飞行计划冲突预探测算法研究[J]. 计算机工程与设计,2006,27(3):430-432.

[9] 陈晓波,宋万忠,杨红雨. 具有多航路点的多机中期冲突探测算法[J]. 计算机工程与设计,2010,31(12):2807-2810.

[10] 查牧言,冯子亮,罗世谦. 适用于多航路的概率型中期冲突探测方法[J]. 计算机应用,2010,30(5):1046-1049.

[11] 崔德光. 空中交通管制自动化中的冲突概率分析[J]. 清华大学学报,2000,40(11):119-122.

[12] 崔德光,王哲鹏. 空中交通管制自动化系统中飞行冲突概率解析算法的应用[J]. 计算机工程与设计,2001,22(5):46-49.

[13] 方保镕,周继东,李医民. 矩阵论[M]. 北京:清华大学出版社,2004.