基于 LSA 和 PLSA 的网页聚类算法研究

Research of Web Page Clustering Algorithm Based on LSA and PLSA

俞 辉 （中国石油大学计通学院 山东东营 257061）
赵玉国 （中国石化胜利油田分公司物探研究院信息技术研究室 山东东营 257022）

摘 要： 文章介绍一种网页聚类算法利用潜在语义分析 LSA（Latent Semantic Analysis）降低词－文档矩阵的秩，在聚类分析中，采用概率潜在语义分析改善聚类精度。首先利用潜在语义分析对词－文档矩阵进行奇异值分解，达到降秩和去噪的目的；然后在聚类分析中，采用概率潜在语义分析设计文档相似度计算函数，实验结果表明该算法的有效性。

关键词： 网页 聚类 潜在语义分析 概率潜在语义分析 相似度

1 引言

随着 Internet 的高速发展，WWW 已经成为当今最大的信息源，其上存储着数以亿计的网页。根据网页内容对其进行聚类分析在信息获取、页面过滤、网页推荐等领域中有着实际应用价值。常用的文本聚类分析利用向量空间模型 VSM（vector space model）将文档表示为词向量，其中每个词对应着一个权值，通过计算文档之间的相似度，聚类相似度大的文档，常用的基于距离的聚类算法有 k－means 算法、k－medoids 算法、BiCH 算法、CURE 算法等。由于文档中出现的词汇量巨大，因此表示文档的向量空间往往是高维的，对其运算的计算量巨大；另外，虽然可以利用词＋权值的形式量化地表示文档，但无法刻画文档的语义，加之文档本身一词多义和多词同义的干扰，造成聚类的准确性不高。

本文利用潜在语义分析 LSA（Latent Semantic Analysis）对词－文档矩阵进行奇异值分解以达到信息过滤和去除噪声的目的，同时矩阵降秩使得向量空间模型中文档的高维表示，投影成在潜在语义空间中的低维表示，缩小了问题的规模。然后利用概率潜在语义分析 PLSA（ProbaBiistic Latent Semantic Analysis）可以将隐类变量 z 与共现数据对——如词汇 w 在文档 d 中的出现——联系成概率统计模型的特点，以此在聚类分析中设计新的文档相似度计算方法，改善聚类精度。

2 相关概念

2.1 潜在语义分析 LSA

潜在语义分析出发点是认为文本中的词与词之间存在某种联系，即存在某种潜在的语义结构，这种潜在的语义结构隐含在文档中词语的上下文使用模式中，通过对词一文档矩阵 A 的奇异值分解计算，并提取 k 个最大的奇异值及其对应的奇异矢量构成新矩阵来近似表示原文档集的词一文档矩阵。词－文档矩阵 A 建立后，利用奇异值分解计算 A 的 k－秩近似矩阵 Ak（k << min(m,n)。经奇异值分解，矩阵 A 可表示为三个矩阵的乘积：

$$A = U\Sigma V^T$$

式中，U 和 V 分别是 A 的奇异值对应的左、右奇异向量矩阵；是标准型；VT 是 V 的转秩；A 的奇异值按递减排列构成对角矩阵 Σk，取 U 和 V 最前面的 k 个列构建 A 的 k－秩近似矩阵

$$A_k = U_k \Sigma k V_k^T$$

式中 U_k 和 V_k 的列向量均为正交向量，假定 A 的秩为 r，则有

$$U^TU = V^TV = I_r$$

可以用 A_k 近似表征原词－文档矩阵 A，其中 U_k 和 V_k 中的行向量分别作为词向量和文档向量。

2.2 概率潜在语义分析 PLSA

给定一个文档集 $D=\{d_1,d_2,\cdots d_n\}$ 和一个词集 $W=\{w_1,w_2,\cdots w_n\}$ 以及一个文档和词的共现矩阵 $A=|a_{ij}|$，其中 a_{ij} 代表不词 W_j 在文档 d_j 中的权值。使用 $Z=\{z_1,z_2,\cdots z_n\}$ 表示潜在语义的集合，k 为指定的一个常数。概率潜在语义分析假设词－文档对之间是条件独立的，并且潜在语义在文档或词上分布也是条件独立的。在上面假设的前提下，可使用下列公式来表示词－文档的条件概率：

$$P(w_j|d_i)=\sum_{k=1}^{k}P(w_j|z_k)P(z_k|d_i)$$

上式中的为潜在语义在词上的分布概率，也可以解释为词对潜在语义的贡献度，通过对排序可以得到潜在语义的一个直观的词的表示。表示文档中的潜在语义分布概率。

概率潜在语义分析使用最大期望 EM（Expectation Maximization）算法对潜在语义模型进行拟合。在使用随机数初始化之后，交替实施 E 步骤和 M 步骤进行迭代计算。在 E 步骤中计算何一个对产生潜在语义 Z_k 的先验概率：

在 M 步骤中，使用下列公式对模型重新估计

$$P(w_j|z_k)=\frac{\sum_{i=1}^{k}a(d_i,w_j)P(z_k|d_i,w_j)}{\sum_{j=1}^{m}\sum_{i=1}^{k}a(d_i,w_j)P(z_k|d_i,w_j)}$$

$$P(z_k|d_i,w_j)=\frac{P(w_j|z_k)P(z_k|d_i)}{\sum_{i=1}^{k}P(w_j|z_i)P(z_i|d_i)}$$

当 L 期望值的增加量小于一个阈值时停止迭代，此时得到一个最优解

$$E(L)=\sum_{j=1}^{m}\sum_{i=1}^{k}a(d_i,w_j)\sum_{i=1}^{k}P(z_i|d_i,w_j)\log[P(w_j|z_k)P(z_k|d_i)]$$

$$P(z_k|d_i)=\frac{\sum_{j=1}^{m}a(d_i,w_j)P(z_k|d_i,w_j)}{ad_j}$$

其中 $a(d_i,w_j)$ 代表词－文档矩阵相应元素的权值。

3 基于 LSA 和 PLSA 的网页聚类算法描述

基于 LSA 和 PLSA 的网页聚类算法由 2 部分组成：数据预处理和聚类分析，下面分别描述这两过程。

3.1 数据预处理

由于网页主要以 html、xml 等格式存储，要抽取网页的文本信息应当对网页进行解析，并分别存为文档。然后基于主题词库，对文档进行分词处理，删除缺乏实际意义的虚词、很少出现的低频词和经常使用的高频词。由于构成文本的词汇数量巨大，因此表示文本的向量空间的维数也相当大，可以达到几万维，必须进行维数压缩的工作。特征提取的方法主要有：文档频率（DocumentFrequency，DF）、信息获取（InformationGAi，IG）、互信息 MutualInformation，MI）、开方拟合检验（CHI，x^2－test）、术语强度（TermStrength，TS）。通过计算词汇的上述任一指标，然后由大到小排序选取固定个数或指标值大于指定阈值的词汇构成特征集。然后对文档特征值进行评估加权，通常权值要考虑来自两方面的贡献，即局部权值和全局权值。在 VSM 模型中局部权值和全局权值有不同的权重取值方法，如 IDF、TFIDF 等。接着构造词－文档矩阵：$A=|a_{ij}|_{mxn}$，其中 a_{ij} 为非负值，表示第 i 个词在第 j 个文档中出现的权重；m，n 分别表示词汇数和文档数。不同的词对应矩阵 A 不同的行，每一个文档则对应矩阵 A 的一列。由于每个词只会出现在少量文档中，故 A 通常为高阶稀疏矩阵。为了将上下文信息考虑进去，需将 a_{ij} 转化为 $\log(a_{ij}+1)$，再除以它的熵（对整行求 plogp），这样预处理能兼顾词的上下文，突出了词在文章中的用文环境. 经过信息熵变换后得到次序化的词－文档矩阵：

$$A'=|a'_{ij}|_{mxn}$$

其中

$$a'_{ij}\frac{\log(a_{ij}+1)}{-\sum_{i=j}\left[\left[\frac{a_{ij}}{\sum_{i=j}a_{ij}}\right]\times\log\left[\frac{a_{ij}}{\sum_{i=j}a_{ij}}\right]\right]}$$

对变换后的文档－词矩阵 A' 进行潜在语义分析，利用奇异值分解计算 A' 的 k－秩近似矩阵 Ak（k << min(m,n)，用 A_k 近似表征原词－文档矩阵，通过奇异值分解和取 k 秩近似矩阵，一方面消减了原词－文档矩阵中包含的"噪声"因素，从而凸现出词和文档之间的语义关系；另一方面使得词、文档向量空间大大缩减，可以提高文本分类的准确率。

3.2 聚类分析

在 PLSA 分析中，可以得到隐含变量 z_k 在文档 d_{ij} 已知的条件下的条件概率 $P(z_k|d_j)$，这样可以构建文档－隐含变量向量 $dl_j=(P_{j,k-1},P_{j,k+1})$ 其中 $P_{j,i}$ 代表文档隐含变量 z_i 在文档 d_j 已知的条件下的条件概率 $P(z_i|d_j)$，向量反映了文档和隐含变量的关系。利用此向量可以计算两文档的相似度，设计相似度计算公式如下：

$$sim(d_i,d_j)=dl_i dl_j/(||dl_i||_2 \bullet ||dl_j||_2)$$

其中 $(dl_i,dl_j)=\sum_{m=1}^{k}P_{i,m}P_{j,m}$，$||dl_i||_2=\sqrt{\sum_{m=1}^{k}P_{i,m}^2}$ 在聚类算法的选择中，本文选用基于距离的 k－medoids 算法，与 k－means 算法比较，该算法选用聚类中位置最靠中心的点做参考点，从而消除了 k－means 算法因采用质心做参考点而导致对孤立点敏感的缺点。同时，该类算法无需事先给出聚类的个数；可以发现非球形的聚类和大小差别很大的聚类；聚类结果与数据输入次序无关并且结果稳定、鲁棒性好。算法在迭代中利用评价函数来选择聚类中心。

3.3 算法步骤

输入值：网页集，降秩矩阵的空间维数 k，阈值 μ，隐含因子数 K

输出值：网页聚类结果 $DC=\{DC_1,DC_2,\cdots DC_l\}$ 和相应的聚类中心 $Cid=\{Cid_1,Cid_2,\cdots,Cid_l\}$ 其中 l 表示聚类数，DC_i 内包含所属的网页

步骤：

Step1：网页解析，将文本信息存储为文档

Step2：文档分词以及词过滤处理，构造文档向量空间

Step3：构造出词－文档矩阵 A

Step4：利用 LSA 对矩阵 A 进行奇异值分解，得到 k 维新的词－文档矩阵 A'

Step5：利用 PLSA 和新的词－文档矩阵 A' 求得每个文档的文档－隐含因子向量 dl

Step6：随机选择 1 个向量初始化聚类 DC1，使 $DC1=\{dl_1\}$，$Cid1=\{dl_1\}$

Step7：对每一个 dl 了，计算其与每个聚类中心点的相似度 $sim(dl,Cid_j)$

Step8：如果 $sim(dl,Cid_j)=max(sim(dl,Cid_j))>\upsilon$，则将 dl 插入 DC_t，并检查更新 Cid_t，否则 dl 将成为新类并成为新类中心。

Step9：如果类中心点不再改变或没有未归类的 dl 则停止，否则重复步骤 7、8

Step10：输出网页聚类结果 $DC=\{DC_1,DC_2,\cdots DC_l\}$ 和相应的聚类中心 $Cid=\{Cid_1,Cid_2,\cdots,Cid_l\}$

4 实验分析

实验网页集抽取自新浪新闻网站 6 个新闻栏目，包括：经济、军事、政治、教育、科技、体育各 50 篇网文，共计 300 篇，分别在不同的控制参数下对算法算法进行比较得到的对比图如下：

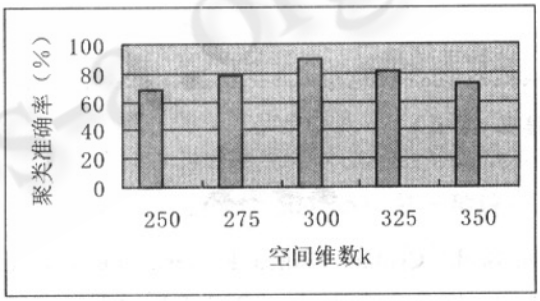

图 1 空间维数对聚类准确率的影响

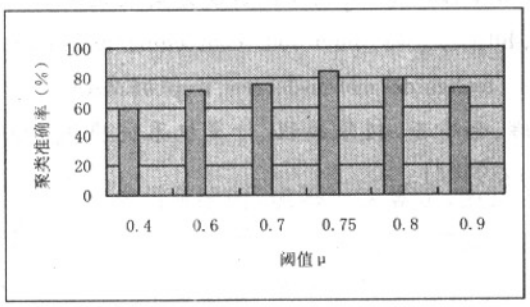

图 2 阈值对聚类准确率的影响

首先可以看出该算法在一定控制参数下具有较高的聚类准确率,同时在图1表明,空间维数k对聚类准确性有一定影响,当k过小时词-文档矩阵压缩过大,无法表述原有语义;当k过大时,降噪效果差,降低了聚类精度。图2表明阈值对聚类准确率影响明显,因为阈值大小直接影响则文档的归属,图中表明当阈值取0.75时聚类精度较高。由于隐含因子数在本算法中只是中间量,因此隐含因子数对聚类准确率的影响不大。

5 结束语

将网页转换为词-文档矩阵,利用潜在语义分析对词-文档矩阵进行奇异值分解以达到信息过滤和去除噪声的目的,同时矩阵降秩使得向量空间模型的维数大为降低,缩小了问题的规模。以k-medodids聚类算法为基础,利用概率潜在语义分析得到的文档-潜在因子的条件概率,计算文档相似度,以此改善聚类精度。实验结果表明此算法的有效性。考虑到提高网页聚类的准确性,下一步应当改进词-文档矩阵权值的计算方法同时应当参考网页的超级链接信息,希望本文工作可以对相关研究提供借鉴。

参考文献

1 S T DumAi, G W Fumas, T K Landauer et al. Using latent semantic analysis to improve information reteieval [C]. In: Proceedings of CHI'88: Conference on Human Factors in Computing. New York: ACM, 1988: 281-285.

2 DumAiS. Improving the retrieval of information fromexternal sources. Behavior Research Methods, Instruments and Computers, 1991, 23(2): 229-236.

3 G. W. 斯图尔特. 矩阵计算引论[M]. 上海:上海科学技术出版社,1980.

4 Hofmann T. Probabilistic Latent Semantic Analysis. Proc. of the 22nd Annual ACM Conference on Research and Development in Information Retrieval. 1999. Berkeley, California: ACM Press.

5 Hofmann T. Latent Semantic Models for Collaborative Filtering. ACM Transactions on Information Systems, 2004, 22(1): 89-115.

6 Y. Zhang, G. Xu, and X. Zhou. A Latent Usage Approach for Clustering Web Transaction and Building User Profile. ADMA2005: 31-42.