LSI 和 kNN 相结合的文本分类模型研究

王天江[1]　叶卫国[1]　卢正鼎[1]　李永平[2]
(1 华中科技大学 计算机科学与技术学院,湖北 武汉 430074;
2 国家药品监督管理局,北京 100810)

摘要: 针对传统文本分类系统的不足,提出了一种基于隐含语义索引的 kNN 的文本分类模型.该方法既充分利用了向量空间模型在表示方法上的巨大优势,又弥补了其忽略语义的不足,具备一定的理论和现实意义.

关　键　词: 文本分类; k 最邻参照法; 隐含语义索引; 奇异值分解
中图分类号: TP301.6　　**文献标识码**: A　　**文章编号**: 1671-4512(2004)04-0059-02

Text classification based on integrating LSI with k-nearnest neighbor

Wang Tianjiang　Ye Weiguo　Lu Zhengding　Li Yongping

Abstract: Because of the deficiency of traditional classification system, the text classification based on integrating k-nearest neighbor with latent semantic indexing was proposed. It took the advantage of abundant expression in Vector Space Model (VSM) and made up the shortage of less semantic information in VSM. The new scheme has significance both in theory and practice.

Key words: text classification; k-nearnest neighbor; latent semantic indexing; singular value decomposition

Wang Tianjiang　Associate Prof.; College of Computer Sci. & Tech., Huazhong Univ. of Sci. & Tech., Wuhan 430074, China.

1　LSI 隐含语义索引

1.1　词-文档矩阵

在 LSI[1] 模型中,文档库表示为 $m \times n$ 的词-文档矩阵

$$\boldsymbol{A}_{m\times n} = [a_{ij}], \tag{1}$$

式中,n 为文档库中的文档数; m 为文档库中包含的所有不同词的个数; a_{ij} 为非负值,表示第 i 个词在第 j 个文档中出现的权重.不同的词对应矩阵 $\boldsymbol{A}$ 不同的一行;每一个文档则对应矩阵 $\boldsymbol{A}$ 的一列.通常 a_{ij} 要考虑来自两方面的贡献,即局部权值 $L(i, j)$ 和全局权值 $C(i)$,它们分别表示第 i 个词在第 j 个文档和整个文档库中的重要程度.这样有

$$a_{ij} = L(i, j)C(i). \tag{2}$$

VSM 模型中局部权值 $L(i, j)$ 和全局权值 $C(i)$ 有不同的权重取值方法[2],如 x^2 (chi-square), IDF, TFIDF. 文献[3]的实验表明,以对数词频法取局部权值和 Entropy 法取全局权值得到的检索效果较好.

1.2　奇异值分解

词-文档矩阵 $\boldsymbol{A}$ 建立后,利用奇异值分解计算 $\boldsymbol{A}$ 的 k-秩近似矩阵 $\boldsymbol{A}_k$ ($k \ll \min(m, n)$).经奇异值分解,矩阵 $\boldsymbol{A}$ 可表示为三个矩阵的乘积:

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}, \tag{3}$$

式中,$\boldsymbol{U}$ 和 $\boldsymbol{V}$ 分别是 $\boldsymbol{A}$ 的奇异值对应的左、右奇异向量矩阵; $\boldsymbol{V}^{\mathrm{T}}$ 是 $\boldsymbol{V}$ 的转秩; $\boldsymbol{A}$ 的奇异值按递减排列构成对角矩阵 $\boldsymbol{\Sigma}$.取 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 最前面的 k 个列构建 $\boldsymbol{A}$ 的 k-秩近似矩阵

$$\boldsymbol{A}_k = \boldsymbol{U}_k\boldsymbol{\Sigma}_k\boldsymbol{V}_k^{\mathrm{T}}, \tag{4}$$

式中 $\boldsymbol{U}_k$ 和 $\boldsymbol{V}_k$ 的列向量均为正交向量.假定 $\boldsymbol{A}$ 的秩为 r,则有

$$\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \boldsymbol{I}_r, \tag{5}$$

收稿日期: 2003-09-22.
作者简介: 王天江(1960-),男,副教授;武汉,华中科技大学计算机科学与技术学院 (430074).
E-mail: wt_jiangmail@yahoo.com.cn
基金项目: 国家高性能计算基金资助项目(00303).

矩阵的分解图如图1所示.

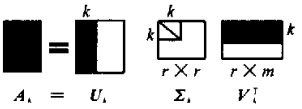

图1 矩阵的SVD分解图

用 A_k 近似表征原词-文档矩阵 A，U_k 和 V_k 中的行向量分别作为词向量和文档向量，在此基础上进行文本分类和其他各种文档处理，这就是隐含语义索引技术. 尽管 LSI 也是用文档中包含的词来表示文档的语义，但 LSI 模型并不把文档中所有的词看作是文档概念的可靠表示. 由于文档中词的多样性很大程度上掩盖了文档的语义结构，LSI 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，从而更加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.

2 LSI 和 kNN 相结合的文本分类模型

kNN 分类算法是一种传统的基于统计的模式识别方法[4]，它将一个文档的所属类别范畴的预测建立在与之最为相似的 k 个文档所属类别的概率分布上. 对一个待分类文档 d，系统在训练集中找到 k 个最相近的邻居，使用这 k 个邻居的类别作为该文档的候选类别. 该文档与 k 个邻居之间的相似度作为候选类别的权重，然后使用预先得到的最优截尾阈值，就可得到该文档的最终分类列表. 本文对 Internet 网上提取出的有关药品广告用语进行检查分类，判定是否属于非法广告用语以及其违法类型.

2.1 训练文档库建模及奇异值分解

由于网上药品广告一般比较短，本文重点讨论对短文档和语句进行匹配分类. 首先建立训练文档库的词-语句矩阵 A. 对该矩阵进行奇异值分解，得到 A 的 k 秩近似矩阵 A_k 以及 V_k. 语句库中的所有语句对应于 V_k 中的一行.

由于药品广告违法类型林林总总，表达形式多样，涉及关键词语也数目繁多. 实际应用中词-语句矩阵 A 是个行列都很庞大的矩阵. 假定有如下5个语句：D_1，该产品安全无毒，根治有效率达90 %；D_2，经国药局认可，该药品能可靠使用；D_3，新产品使用安全方便，无副作用；D_4，该产品配方科学，使用安全，疗效显著，力争达到根治，可靠有效率达99 %；D_5，该减肥药安全可靠，无反弹. 其中 D_1，D_4 和 D_5 属于违法类型中夸大疗效的用语，D_2 和 D_3 属于合法用语.

考虑6个不同的词：T_1，根治；T_2，安全；T_3，有效率；T_4，疗效；T_5，可靠；T_6，科学. 由此可以得到一个归一化的6×5词-语句矩阵

$$A = \begin{vmatrix} 0.5774 & 0 & 0 & 0.4082 & 0 \\ 0.5774 & 0 & 1.0000 & 0.4082 & 0.7071 \\ 0.5774 & 0 & 0 & 0.4082 & 0 \\ 0 & 0 & 0 & 0.4082 & 0 \\ 0 & 1.0000 & 0 & 0.4082 & 0.7071 \\ 0 & 0 & 0 & 0.4082 & 0 \end{vmatrix}$$

A 的秩为4，进行SVD计算后得到矩阵 U，Σ 和 V. 取 $k=3$，得到

$$A_3 = \begin{vmatrix} 0.4971 & -0.0330 & 0.0232 & 0.4867 & -0.0069 \\ 0.6003 & 0.0094 & 0.9933 & 0.3858 & 0.7091 \\ 0.4971 & -0.0330 & 0.0232 & 0.4867 & -0.0069 \\ 0.1801 & 0.0740 & -0.0522 & 0.2320 & 0.0155 \\ -0.0436 & 0.9866 & 0.0094 & 0.4402 & 0.7043 \\ 0.1801 & 0.0740 & -0.0522 & 0.2320 & 0.0155 \end{vmatrix}$$

根据文献[5]中定理有 $\| A - A_3 \|_F / \| A \|_F \approx 0.1876$. 也就是说，将 A 的秩降到3后，矩阵变化了约19 %. 此外，比较 A 和 A_3，注意到 A_3 中的有些元素比原来增强了，有些则减弱了，有些则基本没变. 这种矩阵元素值的此消彼长，使得词和文档之间的语义关系更加清晰.

2.2 待分类文档相似度计算

假设有待分类语句 q，它是一个6维向量，这里直接采用0/1二值法表示. 投影到训练文档库的LSI空间 A_3 后，等效为 V_3 中一行向量. 假设待分类语句为，D：最终达到完全根治艾滋病. 将语句表示为向量形式，即 $q=(1\ 0\ 0\ 0\ 0\ 0)^T$. 投影到 A_3 空间后，有 $\hat{q} = q^T U_k \Sigma_k^{-1} = q^T U_3 \Sigma_3^{-1}$，计算得 $\hat{q} = (0.1575, -0.2301, 0.6317)$. 然后，将 $\hat{q}$ 和语句库中每一个语句向量(这里构造的语句库中语句数目为5)进行匹配，这里选用余弦相似度公式：$\cos\theta_j = \hat{q} s_j / (\| \hat{q} \|_2 \| s_j \|_2)$ $(j=1, 2, \cdots, 5)$，式中 j 表示语句库中第 j 个语句，即 $s_j = V_k^T e_j = V_3^T e_j (j=1, 2, \cdots, 5)$.

最后，分别计算待分类语句 D 和现有语句库中的5个语句 $D_1 \sim D_5$ 的余弦值，得到结果：$\cos\theta_1 = 0.8019$，$\cos\theta_2 = -0.1101$，$\cos\theta_3 = -0.4171$，$\cos\theta_4 = 0.8568$，$\cos\theta_5 = -0.4438$.

(下转第86页)

图2可以看出,对于发射天线×接收天线为5×5的情况,在VBLAST-OFDM系统30 dB的性能水平上,采用 $K=2$ 的VBLAST-OFDM-LCP系统的性能比VBLAST-OFDM系统的性能提高了大约8.5 dB,而 $K=4$ 时提高了大约12 dB.图2还显示,发射天线×接收天线为3×3和5×5的VBLAST-OFDM系统的性能接近,而对于 $K=2$ 和 $K=4$ 的VBLAST-OFDM-LCP系统,3×3比5×5的性能改善略大.

实际上在发射天线数量和接收天线数量相同时,即 $N=M$ 的情况下,Choi Won-Joon等①证明了窄带VBLAST系统的分集增益是1.在宽带信道中,VBLAST-OFDM-LCP系统通过在子载波间进行LCP预编码和交织,将一个符号的信息扩展到多个相互远离的子载波上,因为频率选择性衰落的存在每个子载波上的衰落和信道矩阵有相当大的独立性,LCP预编码的最优分集特性和ML解码使系统获得大小为 K 的频率分集增益.

参 考 文 献

[1] Foschini G J, Gans M J. Layered space-time architecture for wireless communication in a fading environment when using multiple antennas. Bell Labs Syst. Tech. J., 1996, 1: 41~59

[2] Piechocki R J, Fletcher P N, Nix A R, et al. Performance evaluation of BLAST-OFDM enhanced Hiperlan/2 using simulated and measured channel data. Electron. Lett., 2001, 37(18): 1137~1139

[3] Bingham J A C. Multicarrier modulation for data transmission: An idea whose time has come. IEEE Communications Magazine, 1990(5): 5~14

[4] Xin Y, Giannakis G B. High-rate space-time layered OFDM. IEEE Communications Letters, 2002, 6(5): 187~189

[5] Goeckel D L, Ananthaswamy G. On the design of multidimensional signal sets for OFDM systems. IEEE Transactions on Communications, 2002, 50(3): 442~452

(上接第60页)

2.3 各类分值计算

选用 $k=5$ 的 kNN分类方法,即从已知语句库中最多选取5个和待分类语句最相似的训练文本进行比较;同时还需确定判断两语句相似的阈值.不妨假定:对于语句 $\boldsymbol{D}_i$, $\boldsymbol{D}_j$,只有当两语句间夹角 $\cos\theta_{i,j}\geqslant 0.8$ 时,才认为两语句比较相似;即相似度阈值 $\cos\theta=0.8$.

由以上计算结果,只有 $\cos\theta_1$ 和 $\cos\theta_4$ 符合条件.设合法用语类为 c_1,则 $\boldsymbol{D}_2,\boldsymbol{D}_3\in c_1$;夸大疗效的违法类型为 c_2,则 $\boldsymbol{D}_1,\boldsymbol{D}_4,\boldsymbol{D}_5\in c_2$.依据 kNN分类方法,每类的分值为 k 个训练文档中属于该类的文档与待分类文档之间相似之和.这里,符合条件的两语句 $\boldsymbol{D}_1$ 和 $\boldsymbol{D}_4$ 均属于违法类 c_2.即:

$$\text{score}(\vec{d},c_1)=0,$$

$$\text{score}(\vec{d},c_2)=\cos\theta_1+\cos\theta_4=1.6587.$$

2.4 确定待分类文档类型

这里采用最简单的一种判定方法:取阈值 $b=1$,在所有分值超过阈值的类中,判定待分类语句属于最高分值类.显然例子中的待分类语句 $\boldsymbol{D}$:最终达到完全根治艾滋病,属于 c_2 类,即属于夸大疗效类型的违法用语.这也与我们人工判断的结果相吻合.

参 考 文 献

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

[2] Dumais S. Improving the retrieval of information from external sources. Behavior Research Methods, Instruments and Computers, 1991, 23(2): 229~236

[3] Yeung D S, Wang Xizhao. Improving performance of similarity-based clustering by feature weight learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(4): 556~561

[4] Zhang X, Berry M W, Raghavan P. Search schemes for information filtering and retrieval. Information Processing and Management, 2001, 37(2): 313~334

[5] Berry M W, Dumais S T, O'Brien G W. Using linear algebra for intelligent information retrieval. SIAM Review, 1995, 37(4): 573~595

① Choi Won-Joon, Negi R, Cioffi J M. Combined ML and DFE decoding for the V-blast system. in Proc. IEEE ICC-00, New Or-leans, LA, USA, 2000. 18~22

LSI和kNN相结合的文本分类模型研究

作者： 王天江， 叶卫国， 卢正鼎， 李永平
作者单位： 王天江,叶卫国,卢正鼎(华中科技大学计算机科学与技术学院,湖北,武汉,430074)， 李永平(国家药品监督管理局,北京,100810)
刊名： 华中科技大学学报(自然科学版) ISTIC EI PKU
英文刊名： JOURNAL OF HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY(NATURE SCIENCE)
年，卷(期)： 2004,32(4)
被引用次数： 3次

参考文献(5条)

1. Berry M W;Dumais S T;O'Brien G W Using linear algebra for intelligent infomation retrieval[外文期刊] 1995(04)

2. Zhang X;Berry M W;Raghavan P Search schemes for information filtering and retrieval[外文期刊] 2001(02)

3. Yeung D S;Wang Xizhao Improving performance of similarity-based clustering by feature weight learning[外文期刊] 2002

4. DUMAIS S Improving the retrieval of information from external sources 1991(02)

5. Dumais S T;Furnas G W;Landauer T K Using latent semantic analysis to improve information retrieval 1988

引证文献(3条)

1. 张运良.张全 基于句类向量空间模型的自动文本分类研究[期刊论文]-计算机工程 2007(22)

2. 杨清.李方敏 基于潜在语义模型的SVM入侵检测研究[期刊论文]-计算机工程与应用 2007(5)

3. 程传鹏 基于分类的智能信息检索研究与实现[学位论文]硕士 2005