隐含语义索引技术在供求信息分类中的应用

朱学昊[1,2],王儒敬[1]
ZHU Xue-hao[1,2],WANG Ru-jing[1]

1.中国科学院 合肥智能机械研究所,合肥 230031
2.中国科学技术大学 自动化系,合肥 230026
1.Institute of Intelligent Machines,CAS,Hefei 230031,China
2.Department of Automation,University of Science and Technology of China,Hefei 230026,China
E-mail:zxuehao@mail.ustc.edu.cn

ZHU Xue-hao,WANG Ru-jing.Implementation of supply and demand information classification based on latent semantic indexing.Computer Engineering and Applications,2007,43(14):192-194.

Abstract: This paper presents a new implementation of information retrieval and automatic classification.In order to overcome the shortage of traditional methods,an improved classification based on latent semantic indexing is introduced.LSI is a new retrieval model based on Singular Value Decomposition (SVD).Using the algorithm,every term will be either strengthened or weakened. When the latent semantic becomes clearer,it is easy to cut off most of the noisy data at the very beginning.So the accuracy of classification will be improved.

Key words: latent semantic indexing;singular value decomposition;text classification;information retrieval

摘 要:介绍了一种信息抽取和自动分类的新应用,分析了传统分类方法的不足,介绍了一种基于隐含语义索引技术的文本分类改进方案。该技术是一新型的检索模型,它通过奇异值分解,或增强或消减词在文档中的语义影响力,使得文档之间的语义关系更为明晰,从而能容易地剔除掉那些语义关联弱的噪声数据,提高特征值提取精度和最后的分类准确度。

关键词:隐含语义索引;奇异值分解;文本分类;信息抽取
文章编号:1002-8331(2007)14-0192-03 **文献标识码**:A **中图分类号**:TP391

1 研究背景

随着计算机技术的发展,面向农业的信息网站不断涌现。虽然各站点都提供了大量供求信息,但是用户却难以有效地享用这些数据,原因很简单,用户不可能天天都去访问众多网站。这就形成了一种很尴尬的情况,一方面网上存在着大量供求商机,另一方面用户却无法有效使用。能否对这些信息进行统一的抽取和分类,以更全面更合理形式展现在用户面前,这是令人关心的问题。

传统的分类技术大都基于向量空间模型,此模型的前提假设就是特征项之间不相关且分量彼此独立,这就使得它缺乏语义上的约束,不但会因为过早地引入了噪声特征项而削弱最后的分类效果,而且还很难处理歧义和同义的语言现象。

本文将阐述如何对网上供求信息做信息抽取和文本分类,并且针对传统方法的不足,把隐含语义索引技术引入到分类系统中。

2 供求抽取分类系统设计方案

本文抽取分类系统是对各涉农网站所发布的供求信息进行独立抽取和统一分类的应用系统,整个系统设计方案如图1所示。

图1虚线以下的部分是信息抽取模块。该模块针对各个不同结构的站点,分别设计各自不同的包装器(Wrapper)描述各自抽取规则,经过这些包装器的处理后,最终将能形成一个具有上百个站点规模的供求信息量的语料库,并且可每日更新。

图1虚线以上的部分是信息分类模块。因为不同数据源的分类标准往往会彼此冲突,抽取后的数据的分类信息又常会丢失,有些数据本来就被错误分类,所以对供求语料库统一分类的需求变得十分必要。下面将着重介绍这一部分的改进做法。

3 基于语义的去噪声方法

3.1 传统向量空间模型

传统向量空间模型把文档看成是 m 维向量空间中的一个向量,m 为文档中被选中的特征词数量,向量中的元素是对应该特征值的词频。一个文档与某类别的向量间的夹角越小,该文档属于该类别的可能性就越大。

但向量空间模型无法分辨自然语言的语义,它考虑的仅仅是孤立的词频信息,同一类别中所蕴含的词与词之间的语义成分并没有被利用。类别特征不单取决于单个词频也取决于上下文语义对词频的影响。仅仅考虑片面的一个因素往往在处理的

基金项目:国家高技术研究发展计划(863)(the National High-Tech Research and Development Plan of China under Grant No.2003AA118070)。
作者简介:朱学昊(1978-),男,硕士研究生,主要研究方向为网络搜索、数据挖掘;王儒敬(1964-),男,博士,研究员,主要研究方向为智能决策支持系统、数据挖掘。

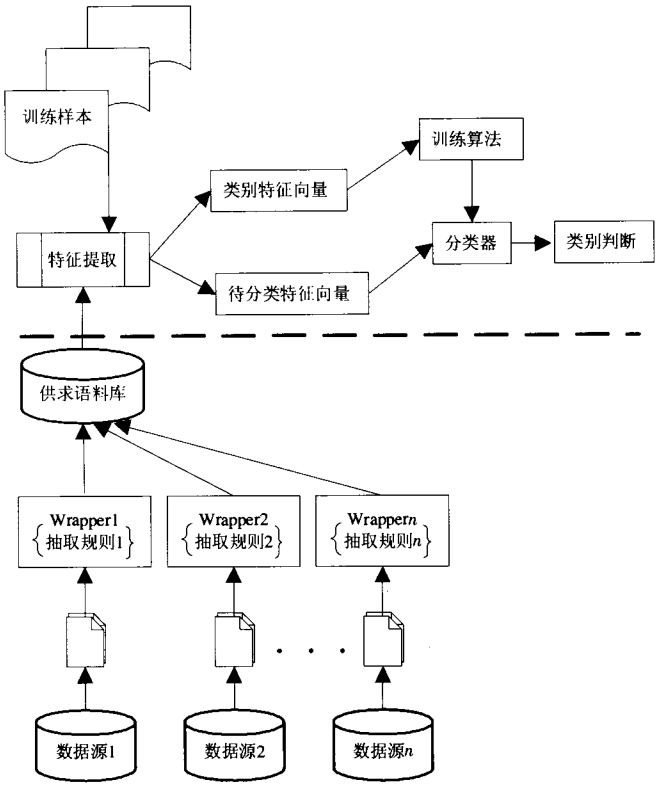

图 1 供求抽取分类系统设计方案

第一个步骤就会引入噪声特征值,这些非特征值对后面的训练和分类是非常不利的。

隐含语义索引(Latent Semantic Indexing)技术正是在这种情况被引入到分类系统中来的,希望能够通过相关的语义计算,强化类别特征值的同时弱化那些非特征值,减少选中噪声特征值的机率,从而有望提高分类的准确率。

3.2 隐含语义索引原理

3.2.1 词-文档矩阵(Term-Document Matrix)的生成

在 LSI 模型中,一个文档库可以表示成一个 $m\times n$ 的词-文档矩阵 $\boldsymbol{A}$。这里,m 表示文档库中包含的所有不同的词的个数,n 表示文档库中的文档数。每一个词对应矩阵 $\boldsymbol{A}$ 的一行,每一个文档对应矩阵 $\boldsymbol{A}$ 的一列。$\boldsymbol{A}=[a_{ij}]$,其中,a_{ij} 为非负值,本文采用权重法对 a_{ij} 进行赋值,即 a_{ij} 的值为第 i 个词汇在第 j 个文本中的权重值。

3.2.2 截断的奇异值分解

利用奇异值分解矩阵 $\boldsymbol{A}$ 后,矩阵 $\boldsymbol{A}$ 就可以表示为三个矩阵的乘积形式:$\boldsymbol{A}=\boldsymbol{U}\sum\boldsymbol{V}'$。其中,矩阵 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 分别是与矩阵 $\boldsymbol{A}$ 的奇异值对应的左、右奇异向量矩阵;矩阵 $\sum$ 是由 $\boldsymbol{A}$ 的奇异值按照递减排列所构成的对角矩阵;$\boldsymbol{V}'$为矩阵 $\boldsymbol{V}$ 的转置。

设矩阵 $\boldsymbol{A}$ 的秩为 r,取 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 最前面的 k 个列构成矩阵 $\boldsymbol{U}_k$ 和矩阵 $\boldsymbol{V}_k$,则可以构建 $\boldsymbol{A}$ 的 k 秩近似矩阵 $\boldsymbol{A}_k$,并且用此矩阵 $\boldsymbol{A}_k$ 来近似表示原词-文档矩阵:$\boldsymbol{A}_k=\boldsymbol{U}_k\sum\boldsymbol{U}_k'$。

其中 $\boldsymbol{U}_k$ 和 $\boldsymbol{V}_k$ 的列向量均为正交向量。其构建步骤如图 2 所示。可知,用简化后的矩阵 $\boldsymbol{A}_k$ 近似表示原词-文档矩阵 $\boldsymbol{A}$ 就是隐含语义索引技术。

虽然 LSI 也是用文档中包含的词来表示文档的语义,但是 LSI 模型并不把文档中的所有的词都看成是文档概念的可靠表示。恰恰相反,LSI 通过奇异值分解和取 k 秩近似矩阵来消减原词-文档矩阵中包含的"噪声"词,强化了词和文档之间的语义关系,凸现类别特征词的权重。总括起来,与传统只基于词

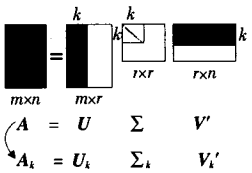

图 2 SVD 图示

频的做法来比较,LSI 的优势表现在:

(1)向量空间中每一维的选取不再只是与是词的简单出现频数和分布有关了,而是隐含了同一类别语义信息。

(2)在特征值选取时,类别"噪声"词的负面影响就会被减弱甚至被消除。

3.3 训练语料库建模及其奇异值分解实例

供求信息一般可分标题部分和正文部分,其标题文字对该分类结果有非常重要的影响,故标题应有较高的权重。设标题词的权值为 3,正文词的权值为 1。假定有如下 5 条供求信息("/"号左边为标题,右边为正文介绍):

D1:供应新鲜葡萄/红提又名红地球葡萄,属欧亚品种。

D2:供应绿葡萄/常年供应新鲜疆绿葡萄。

D3:供应葡萄/新疆葡萄甲天下,尤其以吐鲁番的葡萄最负盛名。

D4:供应红提/果皮中厚,色泽鲜亮,果皮硬脆,新鲜味甜,品质极佳。

D5:供应番茄干/新疆番茄干,颜色鲜艳,营养丰富,风味纯正,用途广泛,自然晾晒。

考虑下面 6 个词:

T1:葡萄

T2:红提

T3:新疆

T4:吐鲁番

T5:新鲜

T6:番茄干

由此可以得出一个没有经过归一化的 6×5 的词-文档矩阵 $\boldsymbol{A}$,矩阵中任意元素 a_{ij} 表示词 i 在文档 j 中的权重:

$$\hat{\boldsymbol{A}}=\begin{bmatrix}4&4&5&0&0\\1&0&0&3&0\\0&1&1&0&1\\0&0&1&0&0\\3&1&0&1&0\\0&0&0&0&4\end{bmatrix}$$

归一化后得到:

$$\boldsymbol{A}=\begin{bmatrix}0.784\,5&0.942\,8&0.962\,3&0&0\\0.196\,1&0&0&0.948\,7&0\\0&0.235\,7&0.192\,5&0&0.242\,5\\0&0&0.192\,5&0&0\\0.588\,3&0.235\,7&0&0.316\,2&0\\0&0&0&0&0.970\,1\end{bmatrix}$$

可知 $\boldsymbol{A}$ 的秩为 5,经过 SVD 计算后得到:

$$\boldsymbol{U}=\begin{bmatrix}-0.924\,0&-0.219\,5&0.124\,0&-0.136\,8&0.055\,0&0.247\,0\\-0.148\,4&0.853\,8&-0.273\,5&-0.398\,4&-0.014\,2&0.123\,5\\-0.153\,3&-0.150\,1&-0.204\,1&-0.292\,2&-0.667\,5&-0.617\,5\\-0.064\,8&-0.044\,8&0.016\,1&-0.300\,1&0.722\,6&-0.617\,5\\-0.310\,1&0.346\,8&-0.081\,8&0.799\,7&0.011\,0&-0.370\,5\\-0.019\,8&-0.279\,3&-0.928\,0&0.087\,7&0.169\,9&0.154\,4\end{bmatrix}$$

$$\Sigma=\begin{bmatrix}1.662\,8 & 0 & 0 & 0 & 0\\ 0 & 1.033 & 0 & 0 & 0\\ 0 & 0 & 0.996\,5 & 0 & 0\\ 0 & 0 & 0 & 0.396\,9 & 0\\ 0 & 0 & 0 & 0 & 0.130\,1\\ 0 & 0 & 0 & 0 & 0\end{bmatrix}$$

$$V=\begin{bmatrix}-0.563\,1 & 0.192\,9 & -0.004\,5 & 0.718\,2 & 0.360\,3\\ -0.589\,6 & -0.155\,4 & 0.049\,7 & -0.023\,6 & -0.790\,7\\ -0.559\,9 & -0.240\,7 & 0.083\,4 & -0.618\,9 & 0.488\,5\\ 0.143\,6 & 0.890\,1 & -0.286\,3 & -0.315\,1 & -0.076\,4\\ -0.033\,9 & -0.297\,4 & -0.953\,2 & 0.035\,9 & 0.022\,8\end{bmatrix}$$

取 $k=3$,得到:

$$A_3=\begin{bmatrix}0.820\,9 & 0.947\,2 & 0.925\,2 & -0.016\,6 & 0.001\,8\\ 0.310\,3 & -0.005\,2 & -0.096\,9 & 0.898\,7 & 0.005\,7\\ 0.114\,6 & 0.164\,3 & 0.163\,1 & -0.043\,2 & 0.248\,7\\ 0.051\,7 & 0.071\,5 & 0.072\,8 & -0.030\,3 & 0.002\,1\\ 0.359\,9 & 0.244\,3 & 0.195\,7 & 0.416\,3 & -0.011\,4\\ -0.033\,0 & 0.018\,3 & 0.010\,7 & 0.012\,7 & 0.968\,4\end{bmatrix}$$

可以看出词-文档矩阵中的有些元素比原来增强了,有些则减弱了,有些基本没有变化,这些元素值的此消彼长,使得词和文档之间得语义关系明确化。根据下面定理,有 $\|A-A_3\|_F/\|A\|_F\approx 0.186\,8$,即当把 A 的秩降为 3 后,矩阵约变化了 18.68%。同样地当取 $k=4$,$\|A-A_4\|_F/\|A\|_F\approx 0.058\,2$,即当把 A 的秩降为 4 后,矩阵只会变化 5.82%。k 取值太小,去除噪声效果并不明显;k 取值太大,则可能会丢失掉不该丢失的特征词。在实际应用中,k 的经验值一般取在 200-300 之间。

分析实例 A_3,词 T6(番茄干)在 D1 中的权值变为更小,因为 D1 所表达的意思与番茄干毫不相关。此类文档中绝大部分与番茄干无关,故词 T6 权值会在此类中急剧下降。同理因为 D5 中有新疆词汇,所以 T6 在同样含有新疆词汇的文档中的权值会发生微弱的提升,但在真实的大样本中,含有新疆词汇的文档同样属于少数,基于这样的提升是无力的。最终词汇 T6(番茄干)在此类中会因为权值小而被当作噪声数据处理。

定理:假设矩阵 A 的 SVD 由方程(3)给出,$r=rank(A)\leqslant \min(m,n)$,且有

$$A_k=\sum_{i=1}^{k}u_i\cdot\sigma_i\cdot v_i^{\mathrm{T}}$$,则

$$\min_{rank(B)=k}\|A-B\|_F^2=\|A-A_k\|_F^2=\sigma_{k+1}^2+\cdots+\sigma_p^2$$

$$\min_{rank(B)=k}\|A-B\|_2=\|A-A_k\|_2=\sigma_{k+1}$$

此定理说明,由 A 的 k 个最大奇异三元组成的 A_k 是和 A 最接近的 k 秩矩阵。

4 分类算法

大量实验和文献表明,K 近邻算法和 SVM 算法是分类效果最好的两个。但 KNN 是一种懒惰的学习方法,平时不训练样本,直到需要分类时才计算分类,速度并不快,基于上述考虑,本系统采用了后一种分类算法。

SVM 理论最初由 V.Vapnik 提出[5],它试图寻找能最大化两类样本之间间隔的最优决策面,以期获得最好的推广性能和最小的训练误差。SVM 在解决小样本,非线性及高维模式识别问题中表现出许多特有的优势,这正是此分类系统所需要的。

5 实验结果

首先把所有涉农的供求信息分成 10 个类别:1 茶叶类,2 瓜果类,3 畜禽类,4 药材类,5 水产品类,6 肉类蛋类,7 粮油蔬菜类,8 林木园艺类,9 农用物资类,10 农业机械类。然后人工搜集了约 4 000 余条分过类的供求信息,力求每类语料库都能涵盖该类所涉及的绝大信息方面。接着从这些语料中提取出 2 000 维的特征向量用于训练分类器。有了所需要的分类器后,就可以对供求语料库里的数据进行分类。图 3 给出了应用 LSI 技术的分类算法与传统分类算法准确率的比较曲线。

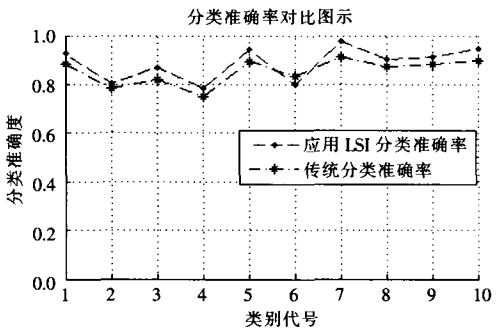

图 3 实验结果对比

由图 3 的实验对比可知,应用了隐含语义索引(LSI)技术的分类器准确率要好于传统基于关键词向量空间模型分析的方法,这是因为每个类别中的噪声词汇通过类别隐含语义的计算在一定程度上减弱了或者排除了。但 LSI 技术尚不能全面理解文本语义,故在某些类别中的表现不如传统方法,这需要语义算法的更深入的研究。整体而言 LSI 技术在克服传统模型中不足的尝试中已取得了很好的效果。

6 结束语

对互联网上现存的农业供求信息进行统一抽取和分类是实际应用中提出的一个重要需求,本文通过分析把隐含语义索引技术引入到传统分类算法中,有效的解决了噪声数据对分类效果的负影响,提高了分类系统的整体准确率。

(收稿日期:2006 年 12 月)

参考文献:

[1] Landauer T K,Dumais S T.Latent semantic analysis and the measurement of knowledge[C]//1st Educational Testing Service Conference on Applications of Natural Language Processing in Assessment and Education,1994.

[2] Liu Tao,Chen Zheng,Zhang Ben-yu,et al.Improving text classification using local latent semantic indexing [C]//Proceedings Fourth IEEE International Conference on Data Mining 2004,ICDM 2004,1-4 Nov 2004:162-169.

[3] 戚涌,徐永红,刘凤玉.基于潜在语义标引的 WEB 文档自动分类[J].计算机工程与应用,2004,40(22):28-31.

[4] 周文,龚礼明,蒋岚.隐含语义检索及中文样本分析实例[J].计算机应用,2004,24(Z1):273-276.

[5] Vapnik V.The nature of statistical learning theory [M].New York:Springer Press,1995.