文本分类中的同义词和关联词聚类算法研究

The Rearch on Synonymy And Association Words Clustering in Text Classification

任红梅 （东营职业技术学院 山东东营 257091）

摘 要：针对基于 VSM 的文本特征空间中存在着大量同义词和关联词的不足，文章结合潜在语义分析和关联规则挖掘以减少信息冗余，改进分类效率。最后对相应的算法进行了描述并实验验证。

关键词：文本分类 潜在语义分析 关联规则挖掘 特征归约

1 引言

当前，随着 Internet 上 Web 信息的增加，文本分类已成为一个日益重要的研究领域。基于文本分类技术有助于解决目前网上信息杂乱的现象，方便用户对信息准确归类和高效检索，是组织和管理数据的有力手段。目前较为常用的文本分类算法有 Bayes、LISF、SVM、KNN、ANN、决策树等。其中大部分分类算法都是基于向量空间模型（VSM），VSM 的基本思想是：文本可以表示成为由独立的概念词组成的向量空间，独立的概念词称为文本特征项，每个特征项是空间一维，维数的权重表示概念的重要程度。向量空间模型的最大优点在于文本表示的可量化，即将文本表示成可度量的向量。作为向量空间的一个点，通过计算向量间的距离决定向量类别的归属。该模型的不足之处在于未能考虑向量模型中各特征项间的相互影响，分类精度不是很理想。事实上构成 VSM 的词语集合并不能完全、准确地反映文本的内容（即：文本的语义）。由于多词同义和语义关联而引起的信息冗余、计算复杂甚至漏判误判的不足非常突出。本文基于自然语言理解技术和数据挖掘技术，对文本特征向量进行同义和关联聚类处理，以实现特征向量的降维和增强独立性的目的。

2 基于 VSM 文本分类方法的分析

大多数分类方法是基于向量空间模型[1]，一个文本 D 可以视为词（或词组）的序列，对于每一词（以下称为向量空间的特征），都根据其在文档中的重要程度赋以一定的权值，从而构成一个向量（$w_1, w_2, \cdots w_n$），其中 w_i 是第 i 个特征的权值，n 是特征总数。一些常用的加权方法包括二进制加权、词频、词频与倒排文档频率之积等。特征向量构造步骤包括：

（1）文档分词处理。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。

（2）特征抽取。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（Information Gain，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。

（3）特征评估加权。其计算方法主要是运用 TFIDF（文本频率及逆文件频率权值）公式，目前存在多种 TFIDF 公式。

当前在文本分类中，对任意两个向量 $X = (x_1, x_2, \cdots x_n)$ 与 $X' = (x'_1, x'_2, \cdots, x'_n)$，存在 3 种最常用的距离度量：欧氏距离、余弦距离和内积，以上 3 种距离度量同样不涉及向量中特征之间的关系。通过上述分析可以发现基于 VSM 模型的文本特征向量虽然有多种抽取方法，但这些方法是根据概率统计的原理设计的，并不考虑特征在语义上的关系，而在自然语言使用中，大量出现同义词和语义关联的现象。如在 IT 技术里“计算机”和“电脑”同义，在司法领域“警察”与“案件”同

时出现的几率特别大。这些同义词和关联词同样出现在特征向量中，一方面增加了特征向量的维数，另一方面降低了特征向量对文档的表达精度。虽然特征抽取可以通过预先设定的阈值来有效降低特征向量的维数，但它不是在基于保证语义精度的前提下，因此常常适得其反。虽然也可以在分词的过程中，使用同义词词典和蕴涵词词典来减少同义词和关联词，这同时也带来词典维护和更新的问题。

3　潜在语义分析和关联规则挖掘

为了解决同义词和关联词对文本分类精度的影响，本文采用潜在语义分析和关联规则挖掘来降低同义词和关联词噪音，提高分类效率和精度。

3.1　潜在语义分析

潜在语义分析 LSA[2,3]（Latent Semantic Analysis）通过引入概念空间来减小同义噪音。LSA 利用词的上下文相关性，即出现在相似上下文中的词，被认为在用法和含义上相近。为实现 LSA 思想，首先构造词－文档矩阵：$A=|a_{ij}|_{m\times n}$，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。不同的词对应矩阵 A 不同的一行，每一个文档则对应矩阵 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)。在 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，如 IDF（逆文件频率权值）、TFIDF 等。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。设第 i 个和第 j 个词分别对应词－文档矩阵的第 i 行和第 j 行，分别记为 $t_i=(a_{i1},a_{i2},\cdots a_{in})$ 和 $t_j=(a_{j1},a_{j2},\cdots a_{jn})$ 其相似度定义为

$$sim(t_i,t_j)=1-\frac{6\sum_{k=1}^{n}(a_{ik}-a_{jk})^2}{n^3-n} \quad (1)$$

在计算相似度之前，通常先将 a_{ij} 转化为 $\log(a_{ij}+1)$，再除以它的熵（对整行求 plogp），这样预处理能将词的上下文考虑进来，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：

$$A'=|a'_{ij}|_{m\times n}$$

$$\text{其中 } a'=\frac{\log(a_{ij}+1)}{-\sum_{i=1}\left(\left(\frac{a_{ij}}{\sum_{i=1}a_{ij}}\right)\times\log\left(\frac{a_{ij}}{\sum_{i=1}a_{ij}}\right)\right)} \quad (2)$$

潜在语义分析的理论基础是矩阵的奇异值分解[4]（Singular Value Decomposition，SVD），奇异值分解是数理统计中常用的方法。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 A_k（$k<<\min(m,n)$）。经奇异值分解，矩阵 A 可表示为三个矩阵的乘积：

$$A=U\Sigma V^T \quad (3)$$

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

$$A_K=U_k\Sigma_k V_k^T \quad (4)$$

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

$$U^TU=V^TV=I_r \quad (5)$$

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。

3.2　关联规则挖掘

关联规则是数据挖掘中的一种主要的挖掘技术[5]，它可以从海量的数据中发现潜在有用的关联或相关关系。设 $I=\{i_1,i_2,\cdots i_m\}$ 是项的集合，记 D 为事务 T 的集合，这里事务 T 是项的集合，并且 $T\in I$。对应每一个事务有唯一的标识，如事务号，记作设 TID。设 X 是一个 I 中项的集合，如果 $X\in T$，那么称事务 T 包含 X。一个关联规则是形如 $X\Rightarrow Y$ 的蕴涵式，这里 $X\subset I, Y\subset I$，并且 $X\cap Y=\Phi$ 规则 $X\Rightarrow Y$ 在事务集 D 中的支持度是事务集中包含 X 和 Y 的事务数与所有事务数之比，记为 support($X\Rightarrow Y$)，即

$$support(X\Rightarrow Y)=|\{T:X\cup Y\subseteq T,T\in D\}|/|D|$$

规则 $X\Rightarrow Y$ 在事务集中的可信度是指包含 X 和 Y 的事务数与包含 X 的事务数之比，记为 confidence($X\Rightarrow Y$)，即 $confidence(X\Rightarrow Y)=|\{T:X\cup Y\subseteq T,T\in D\}|/|\{T:X\subseteq T,T\in D\}|$

给定一个事务集 D,关联规则挖掘问题就是产生支持度和可信度分别大于用户给定的最小支持度和最小可信度的关联规则,常用的算法有 Apriori 算法、FP - Growth 算法。

对于关联词挖掘,设挖掘出的关联规则形如{ $t_i \Rightarrow t_j$,s,c} ,表示了词 t_i 出现在文档中,则词 t_j 出现在同一文档的支持度为 s($0 \leqslant s \leqslant 1$) ,置信度为 c($0 \leqslant c \leqslant 1$)。如果置信度超过事先指定的阈值 σ,则可以认为它们的关联很大,大到即使忽略词 t_j 也不会引起信息损失,这就保证了词的独立性。

4 算法描述及实验结果

4.1 算法描述

通过以上分析构造同义词和关联词聚类算法描述如下

算法 1 :同义词聚类算法.

输入:

(1) 训练文档集,

(3) 概念空间维数 k,

(4) 同义聚类保留特征词个数 N;

输出:

(1) 共 N 个元素的特征词集,

(2) 聚类方案集(决定被聚类的同义语该聚类到那个保留特征词);

步骤

1. 根据训练文档包含词的情况构造出词 - 文档矩阵。

2. 对词 - 文档矩阵进行次序化处理。

3. 使用 SVD 对词 - 文档矩阵进行奇异值分解,得到词 - 文档矩阵的左右奇异向量和奇异值标准型。

4. 保留左奇异向量的前 k 列数据;其它列数据全部清零。

5. 保留奇异值标准型对角线上的前 k 个数据,其它对角线数据全部清零。

6. 保留右奇异向量的前 k 列数据,其它列数据全部清零。

7. 将清零后的左右奇异向量和奇异值标准型相乘,得到新的词 - 文档矩阵。

8. 对新的词 - 文档矩阵进行次序化处理。

9. 当词个数超过 N 时做步骤 10 ~ 14 ,进行同义聚类。

10. 在特征词集合中查找相似度最大的两个特征词。

11. 在特征词集中删除相似度最大的特征词对中的任一个。

12. 在聚类方案集中查找含有相似度最大的特征词对中任一特征词的聚类方案。

13. 如找到,则将另一个特征词加入到聚类方案的聚类特征词集合中去。

14. 如匹配不到聚类方案,则以这两个特征词构造一个聚类方案,并放入聚类方案集中去。

算法 2 关联词聚类算法.

输入:

(1) 训练文档集,

(2) 同义聚类后的特征词集,

(3) 关联聚类阈值 σ,

(4) 同义聚类后的聚类方案集;

输出:

(1) 关联聚类后的特征词集,

(2) 关联聚类后的聚类方案集;

步骤

1. 利用 Aprior 算法求得所有置信度大于 σ 的单维关联。

2. 对每一条置信度大于 σ 的单维关联规则做步骤 3 ~ 6 ,进行关联聚类。

3. 在特征词集中删除关联规则右部的特征词。

4. 在聚类方案集中查找含有关联规则任一边特征词的聚类方案。

5. 如找到,则将另一特征词加入到聚类方案的聚类特征词集中去。

6. 如匹配不到聚类方案,则以关联规则左右部的两个特征词构造一个聚类方案,并放入约方案集中去。

4.2 实验结果

选择中国石油大学(华东)新闻网(http//:news.upc.edu.cn)中的 8 个栏目,每个栏目随机选择 20 条新闻语料构成训练文档集。考虑日常使用的汉语词汇不过 3 千多条而所选语料的每篇文章最多 1 千字左右,而绝大多数文章仅有 4、5 百字,因此设定同义特征词集元素数 N = 1200;为了控制程序运行时间设定概

(下转第 31 页)

(上接第 26 页)

念空间维数 k =20,关联归并的支持度 s =1%,置信度 c =1.5%;考虑算法中有较多的矩阵运算,算法采用 Matlab 编程实现。最后生成 1200 个同义特征词,每个同义特征词对应 0 ~6 个同义词;生成 1556 个关联特征词,每个关联特征词对应 1 ~13 个关联词。经过人工检查,像"教师"与"老师"、"教学"与"授课"等同义词以及"石油"与"加工"、"油气"与"储运"等关联词都可以成功提取,结果令人满意。

5 结语

目前对语义精确理解还很难做到,潜在语义理解却是实际可行的。结合自然语言理解技术和数据挖掘技术,对文本特征向量空间进行同义聚类和关联规约可以有效地减少信息冗余,有利于文本快速准确分类,提高分类精度。

参考文献

1 叶新明、徐进鸿,中文文献自动分类研究[J],情报学报,1992;(5).

2 S T Dumais, 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.

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

4 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.

5 史忠植,知识发现[M],北京:清华大学出版社,2002.