一种基于本体论的文本特征选取方法*)

林东文　白清源　谢丽聪　谢伙生　张　莹
(福州大学数学与计算机科学学院　福州 350002)

摘　要　针对文本特征向量高维数的问题，给出了一种基于本体论的文本特征选取方法。通过由专业领域本体所建立的概念树，把文本的特征项映射到概念，同时进行特征项频度到概念频度的转换，使得选取得到的特征概念能够很好表征文本的内容。实验结果表明，与未进行特征概念选取相比，采用此方法选取得到的特征概念能够在尽可能减少对文本分类精度的影响下，达到降低特征维数的目的。

关键词　本体，文本特征，文本分类，特征选取

A Ontology-based Document Feature Extraction

LIN Dong-Wen　BAI Qing-Yuan　XIE Li-Cong　XIE Huo-Sheng　ZHANG Ying
(College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350002)

Abstract　To effectively reduce the dimension of document vectors, we introduce a novel method employing domain ontology to extract feature concept. For all document categories, all raw words in each category are mapped to concepts in their relative concept tree derived from the domain ontology. At the same time the frequency of raw words is transformed into the frequency of concepts. Experimental results show that this method can effectively reduce the dimension of document vectors without loss of categorization accuracy, compared with traditional document vectors.

Keywords　Domain ontology, Document feature, Text classification, Feature selection

1　引言

目前，文本分类领域遇到的重要问题是海量的文本信息如何用数量有限的特征集合表示。在许多文本数据集中，只有很小部分的特征集合在文本分类中是有用的，使用所有的特征则会降低文本分类的性能。因此，如何有效地降低特征向量维数成为文本分类研究的重要问题。停词列表[1]和词干提取[2]成为该问题上的早期解决方法。

近几年国内外出现了一些基于本体的特征提取算法[3~5]，它们能够在有限度降低分类精度的基础上减少特征空间的维数。Hotho, Staab, Madche 等人[3]提出了使用本体字典(wordnet)得到特征概念子空间并应用到本体聚类的特征提取阶段中。实验结果表明该方法优于使用 if/idf 方法的聚类算法。Bill B. Wang, R. I. McKay 等人[4]提出了在已构建的本体库上使用启发式函数得到优化的特征概念空间，用以解决测试样本特征词不出现在训练样本的情况。实验结果表明该方法优于使用特征词和未优化的特征概念集合。

此外，中国科学院计算所的 Kai Zhang, Jian Sun, Bin Wang 等人[6]提出了利用 wordnet 解决同时出现在两类文本中的特征词选取问题。实验结果表明该方法适用于小规模的文本分类，对主题一般化的两类文档分类错误率较高。

本文提出了一种基于本体论的文本特征概念选取方法，主要用于与某专业领域相关的文本分类的特征提取阶段。利用本体论知识减少特征向量空间的维度，改善特征词对文本的贡献程度，达到提高分类器分类精度的目的。本文的组织结构如下：第 2 节介绍了领域本体的概念及本文所采用本体的主要来源；第 3 节对本文提出的特征概念选取法进行详细的论述；第 4 节是对实验结果的分析和比较。最后对文本的研究工作和下一步研究方向进行总结和展望。

2　领域本体与 UMLS

在计算机科学方面，通常使用本体对外在世界的对象进行形式化描述，用符号表示对象的属性及其它们之间的关系。我们将利用本体知识得到代表一文本的概念集合。

本文所采用的领域本体框架定义如下：

定义 1(领域本体)　一个核心本体为一符号系统 $O=\{L,C,H,ROOT,F\}$。

L：词典，即词项集合。

C：概念集合。

H：概念之间在语义上直接的、无循环的、传递的、自反的层次结构。比如：H(RNA 病毒，病毒)表示 RNA 病毒是病毒的子概念。

$ROOT$：概念集合 C 中的最顶端概念。C 中的其他概念都是 $ROOT$ 的子概念。

F：映射函数：$2^L \rightarrow 2^C$。表示词典 L 中的词项集合对应概念集合 C 中的概念集合。一般来说，多个词项对应一个概念，或者一个词项对应多个概念。

领域本体包括了所有与该领域中出现的概念和关系。本文所采用的本体知识主要来自统一的医学语言系统(UMLS)。它由美国国立医学图书馆自 1986 年开始研制，该系统由超级叙词表、语义网络、信息资源及专用词典组成。该系统定义了概念间的多种关系类型。ISA 是最主要的关系之一。本文将使用该关系为每个训练文本集分别建立相应的概念层次结构。

*)由福州大学科技发展基金(2005-XQ-13、2006-XQ-22、XRC-0511)、福建省教育厅(JB06023)资助。**林东文**　硕士研究生，主要研究方向：特征提取、本体；**白清源**　副教授，研究方向：数据库技术和数据挖掘。

3 建立文本的优化概念向量空间

本文主要解决的是在文本分类中的特征提取问题。针对狭义的文本分类即对与某专业领域(如医学)相关的文本进行分类。建立文本的概念特征向量分为以下三个步骤：

1. 通过查询 UMLS 为每个训练文本集分别建立以类别名为根的概念树。

2. 找出每个文本中所有的特征项及其出现次数。通过查找文本所在类别的概念树，把特征项映射到概念，同时根据映射关系计算概念的频度，得到代表文本的概念向量。

3. 对于步骤 2 所采用的 TTCmap 算法的具体描述。

3.1 为每个训练文本集分别建立以类别名为根的概念树

本文主要考虑的是狭义文本分类即与某专业领域相关的文本分类，采用的本体知识来自统一的医学语言系统(UMLS)。假定每个文本只属于医学领域的一分支科目，比如病毒学，免疫学等，也就是每个文本只在一类别中出现。

设待分类的文本集合为 D，共有 n 个分类，分别为 $D1$ 到 Dn；相应的类别名为 $N1$ 到 Nn；每个分类中的文本数为 M。根据每个训练文本集 $Di(i=1\cdots n)$ 的主题(类别名)，通过 UMLS 系统提供的 API 得到以一类别名 Ni 为根的概念树 Ti。概念树中概念的关系目前只考虑 ISA 的关系。这样，就为训练样本集中的每个文本集建立了一一对应的概念树，也就是 Di 到 $Ti(i=1\cdots n)$ 的一一对应。

3.2 特征项映射到概念的具体分析

目前采用最广泛的文本表示方法为特征向量空间。该方法的主要缺点是特征向量的维度过高，通常超过 10000 个特征项。实际上，向量空间中许多完全不同的特征项所代表的概念是相同的。对于本文中所要解决的狭义的文本分类，由于领域中的概念通常为复合词，因此会出现多个特征项共同对应一个概念的情况。

把特征项映射到概念的过程主要是在得到一文本 d 中出现的所有特征项之后进行。映射的过程主要根据 3.1 节中得到的文本所在类别对应的概念树。由于概念由多个词组成，因此在去除停用词并提取词干后得到的特征项实际上很难在概念树中找到完全匹配的概念。考虑到专业领域的概念通常是由复合词组成，所以在进行特征项到概念的映射过程中，采用最大匹配的多个特征项共同映射一概念的方法。也就是，构成一概念的所有词是否都为文本的特征项，如果所有词都在一文本的特征向量空间中出现，就可以得到这些特征项到该概念的映射关系。例如：catalyz，rna 这两个特征项共同映射到概念(Catalytic RNA)(C0080125)(注：C0080125 为 UMLS 中该概念的 ID 号)。目前考虑只要一概念中半数以上的词在文本的特征向量空间中出现，就可以建立这些特征项到概念的映射关系。

实际中，把多个特征项映射到概念时会出现两类情况。一类是不出现多个特征项交叉映射到多个概念的情况(1)。另一类是出现多个特征项交叉映射到多个概念的情况(2)。具体符号描述如下：

$TS=\{t\in T|c1\in F(t)\}$　$TS2=\{t\in T|c2\in F(t)|$

其中 TS 称为特征项集合，表示在文本 d 中映射到概念 c 的多个特征项组成的集合。$TS1$ 表示与概念 $c1$ 相对应的特征项集合。$TS2$ 表示与概念 $c2$ 相对应的特征项集合。

(1) $TS1\cap TS2=\phi$　(1)

表示两特征项集合中不存在相同的特征项。

(2) $TS1\cap TS2\neq\phi$　(2)

表示两特征项集合中存在相同的特征项。

例如：特征项集合 $TS1=\{rna, small, Interfer\}$ 映射到概念 $c1(RNA, Small\ Interfering)$(C1099354)，另一个特征项集合 $TS2=\{small, nuclear, rna\}$ 映射到概念 $c2(Small\ Nuclear\ RNA)$(C0035709)，这符合(2)。

对于这两种映射情况，在映射得到的概念频度计算上需要分别考虑。对于情况(1)，由于一概念的特征项集合 TS 同其他所有概念的特征项集合无交集，那么可以认为特征项集合 TS 中的所有特征词都是在一文本中唯一出现的，根据 TS 中所有特征词的共现情况取 TS 中频度最小的特征项频度作为该概念的频度，见式(3)。同情况(1)相比，情况(2)的概念频度计算相对复杂些，主要的考虑原则是从一概念的特征项集合 TS 中找出最能代表该概念，也就是和其他概念区分度最大的特征项，把它作为该概念的代表，相应地，该概念的频度就是该区分度最大的特征项的频度，见式(4)。

那么，对于情况(1)，映射得到概念 C 的频度计算公式如下：

$$cf1(d,c)=\min\{tf(d,TS)\}$$
$$TS=\{t\in T|c\in F(t)\}\quad(3)$$

其中 $cf1(d,c)$ 表示概念 c 在文本 d 中的频率。min 表示取特征项集合中频度最小的特征项频度作为概念 c 的频度。

然而，对于情况(2)，考虑保留多个特征项映射得到的多个概念。该情况下概念 C 的频度计算公式如下：

$$cf2(d,c)=tf(d,\min(exist(d,TSi)))\quad(4)$$
$$exist(d,TSi)=\{t|t\in TSi\ and\ t\in TSj\}\ (j=1\cdots n, j\neq i)$$

其中 $exist(d,TSi)$ 表示在概念 Ci 的 TS 中出现、同时也在其他概念 Cj 的 TS 中出现的特征项集合，并记录每个特征项在其他概念 Cj 的 TS 中的出现次数。$\min(exist(d,TSi))$ 表示选取在 $exist(d,TSi)$ 中出现次数最小的特征项。$cf2(d,c)$ 表示把出现次数最小的特征项在文本 d 中的频率作为概念 C 的频度。

3.3 TTCmap 算法的具体描述

那么，由 3.2 节中的问题分析可以得到映射算法 TTCmap 的具体描述：

输入：文本 d 中出现的特征词集合 $T=\{t1,t2,\cdots,tn\}$。
输出：概念集合 C。
初始状态：概念集合 $C=\phi$，临时概念集合 $TempC=\phi$。
扫描 T 中的所有特征项 $ti(i=1\cdots n)$，找出概念树中同每个 ti 模糊匹配的所有概念，把这些概念存入临时概念集合 $TempC$。
在 $TempC$ 中删除所有只出现一次的概念(根结点上的概念除外)；
找出与每个概念相映射的特征项集合 $TSi(i=1\cdots n)$；
While(TSi)
{
　如果 TSi 属于情况(1)，
　　则概念 Ci 的频度计算采用公式(3)；
　如果 TSi 属于情况(2)，
　　则概念 Ci 的频度计算采用公式(4)；
　$TempC=Ci\cup TempC$;
}
Return $C=TempC$;

该算法中提到保留根结点上的概念，主要考虑到根结点的概念通常由单个词构成，在所在类别的文本中经常出现，所以予以保留。

通过 TTCmap 算法可以得到代表一文本 d 的概念向量。由于多个特征项映射一概念，那么概念向量的维度肯定小于原先特征向量的维度，同时概念相对于特征项更能表征文本的内容。以上所提到的特征项频度和概念频度指特征项或概念在一文本 d 中的出现次数。在接下来的实验中我们采用 KNN 分类器作为评价该算法的分类算法。在第 4 节中可以看到概念特征向量下的 KNN 分类器的分类精度比一般特征向量的分类精度高，也能够明显降低向量空间的维度，减少

KNN 分类器的计算量。

4 实验结果的分析和比较

实验过程中使用类似 tfidf 方法来计算概念 c 的频度。把原先考虑的特征项 t 出现次数改为概念 c 的出现次数。概念 c 在文本 d 中的频度计算公式如下：

$$\text{tfidf}(d,c)=\log(lf(d,c)+1)*\log\left(\frac{|D|}{df(c)}\right)$$

一文本 $\vec{d}$ 的特征概念向量(CF)表示为

$$\vec{d}=(cf_1,cf_2,cf_3,\cdots,cf_n)\ (cf_i=\text{tfidf}(d,ci))$$

我们从医学数据库(PUBMED)中选取了 5 份杂志的内容作为实验的训练样本和测试样本(如表 1)。从表 1 中可以看出这些样本的主题是相互独立的。从每份杂志中分别选取了 125 篇论文(包括论文的标题和摘要)，其中 100 篇作为训练文本，25 篇作为测试文本。对所有类别中的测试文本和训练文本都已经过人工标定类别，以数字形式表示各文本，所有文本均进行了预处理包括去除停用词和提取词干。实验中采用 KNN 分类器来评估文中提出的特征概念选取方法的效果。将未进行映射的一般特征向量同概念特征向量进行比较。具体比较结果如表 2。实验中采用的评价参数如下：

$$\text{每个分类的准确率}=\frac{\text{该分类的正确文本数}}{\text{该分类的实际文本数}}$$

$$\text{总分类的准确率}=\frac{\text{分类的正确文本总数}}{\text{分类的实际文本总数}}$$

表 1 训练文本数据和测试文本数据描述

杂志名称	类别名称	出版年份
RNA	RNA	2006
Genetics	Genetics	2006
Texas Heart Institute Journal	Heart	2005 2006
Clinical Microbiology Reviews	Microbiology	2002 2003 2004 2005 2006
Epilepsy Currents	Epilepsy	2004 2005 2006 2007

表 2 训练样本的分类准确率

类别名称	一般特征向量	概念特征向量
RNA	64%	88%
Heart	44%	80%
Genetics	56%	92%
Microbiology	68%	76%
Epilepsy	76%	68%
总分类的准确率	61.6%	80.8%

表 2 列出了使用一般特征向量和使用概念特征向量时，KNN 分类器的分类准确率。可以看出在五类文本集中，大部分文本集的概念特征向量相对于一般特征向量得到的分类准确率更高些，其中 Epilepsy 类别的分类准确率反而有所下降。主要是由于该类别对应的概念集合中的概念数量相对较少，KNN 分类器在进行概念频度选取时得到的该类别概念较少。表 3 列出了训练样本数量不同时得到的特征词个数，可以看出通过 TTCmap 算法得到的概念数量明显低于特征项的数量。这反映了采用该映射算法能够达到降低向量空间维度的目的。

结论和今后的研究工作 以上实验结果表明：基于一般特征向量的文本表示法存在大量的无意义词项，在训练文本集合小时，分类精度不高。本文提出的基于本体的概念特征向量的文本表示方法，通过把词项映射到概念并进行词频到概念频度的转换计算，得到的概念特征向量能够有效提高训练文本集合很小时 KNN 分类器的分类精度。此外，由于本体中的概念数量有限，可以预见当训练文本集合很大时，词项数目有上万，如果采用把本体中的概念代替词项，则可以用有限的概念来表示训练文本集，达到降维的目的。

表 3 有关训练文本数据的统计信息

训练样本大小	特征词	概念
750	5867	1530
600	3895	1271
500	3211	1063
300	2426	985

然而，我们考虑到本文的文本范围与某专业领域相关，领域中各分支之间的本体概念是否会出现重叠现象，即多学科交叉的情况。这种情况的出现是否会对从特征项映射到概念的过程产生一定的影响。这将是我们未来的研究工作。

参 考 文 献

1 Fox C. Lexical Analysis and Stoplists. In Information Retrieval: Data Structure & Algorithms. In: Frakes W B, Baeze-Yates R, eds. P T R Prentice Hall, 1992. 102～130

2 Frakes W B. Stemming Algorithms. In Information Retrieval: Data Structure & Algorithms. In: Frakes W B, Baeza-Yates R, eds. T P R Prentice Hall, 1992. 131～160

3 Hotho A, Staab S, Maedche A. Ontology-based Text Clustering. IJCAI'01-Workshop Text Learning: Beyond Supervision. Seattle, USA, 2001

4 Bill b, McKay R, Abbass H A, Michael B. A Comparative Study for Domain Ontology Guided Feature Extraction. In: Proc. of The Twenty-Fifth Australian Computer Science Conference. Conferences in Research and Practice in Information Technology, 2003, 16

5 Hotho A, Staab S, Stumme G. Wordnet improves Text Document Clustering In: Proc. of the SIGIR 2003 Semantic Web Workshop, 2003

6 Zhang Kai, Sun Jian, Wang Bin. A Wordnet-based Approach to Feature Selection in Text Categorization Intelligent information processing II table of contents, 2004

(上接第 138 页)

3 Picard R W, Vyzas E, Healey J. Toward Machine Emotional Intelligence: Analysis of Affective Physiological State. IEEE Transactions Pattern Analysis and Machine Intelligence, 2001, 23(10) 1175～1191

4 Wagner J, Kim J, André E. From Physiological Signals to Emotions: Implementing and Comparing Selected Methods for Feature Extraction and Classification. In: IEEE International Conference on Multimedia&Expo (ICME 2005), 2005

5 边肇祺，张学工. 模式识别. 第二版. 北京：清华出版社，2000

6 Kennedy J, Eberhart R. Particle swarm optimization. Proceedings of 1995 IEEE International Conference on Neural Networks, 1995, 4: 1942～1948

7 Kennedy J, Eberhart R C. A discrete binary version of the particle swarm algorithm. In: Proc. of the 1997 Conf on Systems, Man, and Cybernetics. Piscataway: IEEE Press, 1997. 4104～4108

8 Mohan C K, Al-kazemi B. Discrete Particle Swarm Optimization. In: Proc. Workshop on Particle Swarm Optimization, Indianapolis, Purdue School of Engineering and Technology. IUPUI, 2001

9 Shi Y, Eberhart R C. A modified swarm optimizer. In: IEEE International Conference of Evolutionary Computation. Alaska: Anchorage, IEEE Press, May 1998

10 http://mm-werkstatt. informatik. uni-augsburg. de/team_details. php? id=32