# 基于知识本体的文本分类技术及其应用研究*)

李志国[1,2] 钟 将[1] 冯 永[1] 叶春晓[1]

(重庆大学计算机学院 重庆 400044)[1] (上海宝信软件西南研发中心 重庆 400041)[2]

**摘 要** 文本分类技术是知识管理系统实现知识有效组织、存储和检索的重要手段,而基于词向量空间模型的文本分类没有考虑知识管理系统的特点,从而也不能满足知识管理系统中多分类的需要。论文提出了一种新的基于知识本体的文本分类算法,该方法利用知识管理系统中的本体集,实现了多概念粒度分类,实验表明该方法具备良好的分类性能。

**关键词** 知识管理系统,文本分类,本体,多粒度分类

## Study on the Text Classification Algorithm Based on Knowledge Ontology

LI Zhi-Guo[1,2] ZHONG Jiang[1] FENG Yong[1] YE Chun-Xiao[1]

(College of Computer Science, Chongqing University, Chongqing 400044)[1]

(Southwest Research Center of Shanghai Baosight Software Corporation, Chongqing 400041)[2]

**Abstract** In the knowledge management system, text classification technology is an important means of organizing, storage, and retrieval for the knowledge. The traditional text classiication technology is difficult to meet the needs of the knowledge managemetn system. In this paper, a new ontology-based text classification algorithm is presented, which takes full advantage of the knowledge management system featrues and implements the multi-granularity classification for the concepts. The experiments show that the method has a good classification performance.

**Keywords** Knowledge management system, Text classification, Ontology, Multi-granularity classification

## 1 引言

随着网络技术的迅速发展,企业内部的数据规模呈指数增长,如何从海量信息中搜索、过滤、管理这些数据资源成为十分重要的问题。其中以知识管理系统为中心的数据组织和管理的模式逐渐成为企业的首要解决方案[1]。

所谓知识管理就是企业对其所拥有的知识资源进行管理的过程,它运用集体的智慧提高应变能力和创新能力,为企业实现显性知识和隐性知识的共享提供了新的途径。在信息系统的支持下,知识管理通过创造一种便于每位成员获取、共享和使用组织内、外部知识的环境,支持把知识应用到组织提供的产品和服务中去,最终提高企业创新能力和对市场反应速度。

文本分类(Text Classification)是指由计算机自动提取文本的特征,依据一定的算法,将文本按内容或属性归到一个或多个类别的过程。因此文本分类技术有助于知识的组织和管理,进而建立合理的知识分类库,对于提高知识检索效率十分有效。目前已有许多机器学习方法应用到文本分类中,如Vapnik 提出的支持向量机(SVM)[2]、K 近邻(KNN)分类器[3]、Generalized Instance Set 的方法[4]等。这些分类算法基于文档的向量空间表示模型,然后在每个类别的训练文本集合的基础上训练出一个分类器,最后通过分类器将文本分类。

在知识管理系统中的知识获取、存储和检索以及共享等关键处理过程中都需要使用到文本分类技术。例如在知识获取阶段,就需要判断那些文档和知识是当前企业感兴趣的知识,需要将这些知识归为那一类,并提供给哪些用户。

然而现有的文本分类方法通常考虑的是特征词的词频,没有考虑词之间的语义关系果,以及其具体的应用领域,因此这些文本分类技术还不能完全满足知识管理的需要。本文根据知识管理系统的特点提出了一种新的基于本体的文本分类技术。

## 2 面向企业知识管理系统的文本分类

由于知识管理系统通常是面向行业应用,因此企业知识管理系统中存在有别于一般文本检索系统的特点:

1)由于知识管理系统通常是面向特定行业和领域的,因此可以通过行业专家或者行业知识建立一个受控的术语表(词汇表)来限定和描述所关心的知识以及知识之间的关系。因此,文本分类系统可以充分利用这些词汇表实现高效和精确的文档分类和检索。

2)知识通常具备一定的模糊性,存在多分类的特性,即同一个知识可能会被划分到多个知识类别中。因此,在文本分类时需要根据一定度量方法将文本文件划分到多个类别中,以提高用户在检索和利用知识过程中的查全率。

3)知识之间具备复杂的关联关系,用户在使用某个知识时,可能需要了解与此相关联的其它知识。那么这就为文档分类系统带来较大的挑战。

4)同时知识管理中的文本分类是一个多粒度的分类。例如对于一个 IT 产品销售企业,一个属于市场方面的知识,同时它可能是关于笔记本电脑的市场方面的知识。因此需要实

*)基金项目:浦东新区科技发展基金 PKK2005-07;国家发改委科学研究计划项目 2005-2137。李志国 博士研究生,主要研究方向:知识管理与知识发现。

现多概念粒度的文本分类技术的支持。

为适应知识管理系统的这些特点，本文提出了一种基于本体的文本分类技术。该方法利用企业知识管理系统中知识本体和受控关键词表，基于概念之间的相似度来实现文本的精确的查询和检索。

## 3 知识本体中概念间的相似度计算

知识管理的核心问题是实现组织内不同用户之间知识共享从而推动知识创新。实现知识共享的前提是组织内部的用户需要采用统一的知识表示和描述的方法。

传统的知识管理系统则采用概念树的方法，其实质就是借鉴图书管理中按照学科门类组织不同书籍的方式来组织和管理知识库。尽管该方法具有直观，易于实现的优点，然而它只能表示概念之间的从属关系，包含的语义信息较少。因此研究者逐渐开始转向具有丰富语义信息和推理能力的本体论方法来组织和描述知识[5]。

本体可以通过定义精确的共享术语，提供某一特定领域可重用的知识。在知识管理系统中，本体的应用主要体现在如下几个方面[6]：(1)通过提供标注文档内容的语义信息，大大提高检索的查全率和查准率；(2)本体提供一种组织架构对多个信息源进行信息集成，从而有利于数据、知识和模型交换；(3)通过对信息内容的约束确保一致性和正确性；(4)创建可相互交换和可重用模型库；(5)支持从文档集中提取附加知识。

**定义 1** 概念树可以表示为一个二元组的集合 $CT=(C,R)$，其中 $CT$ 表示概念树，$C$ 表示概念树中的概念，$R$ 表示概念之间的关系，由于概念树只描述概念的从属关系，例如概念“篮球运动”是“体育运动”的子概念，就表现为树上的子节点与父节点之间的关系。

**定义 2** 本体可以表示为一个五元组的集合 $O=(C,W,R_w,R_c,T)$，其中 $C$ 为概念集合，是指客观世界的实体，或者实体在思维世界的反映；$W$ 为词汇是语言世界的符号，这种清晰的符号为人与机器之间提供了共同理解的概念基础；$R_w$ 为词汇到概念的映射；$R_c$ 为概念之间的关系的集合；$T$ 为本体中顶层概念集合。

词汇到概念之间的关系是简单的一对多的关系，例如概念“计算机”对应的词汇有“PC”，“电脑”，“Computer”等。

本体集中的概念之间的关系存在多种关系，从知识管理的实际需求出发，本文从语义上讲，基本的关系共有 5 种，如表 1 所示。

表 1 本体中概念间的基本关系

| 关系名 | 关系描述 |
|---|---|
| Part-of | 表达概念之间部分与整体的关系。 |
| Kind-of | 表达概念之间的继承关系，类似于面向对象中的父类与子类之间的关系。 |
| Instance-of | 表达概念的实例与概念之间的关系，类似于面向对象中的对象和类之间的关系。 |
| Attribute-of | 表达某个概念是另一个概念的属性。如“价格”是桌子的一个属性。 |
| Link-of | 表示两个概念之间存在关联关系，这种关联关系主要是满足用户自定义各种关系，可以看成是上述关系的一般形式。 |

由此可见，本体中概念之间的关系比概念树的关系更丰富和灵活，表示语义信息更丰富，因此目前的知识管理系统的研究者和开发者试图使用本体来描述知识以及知识之间的关系。

显然对于从属于同一顶层概念的概念之间存在某种联系，并且这种联系存在强弱之分。为了定量描述概念之间联系的强度，做以下的定义。

**定义 3** 相邻概念，是两个概念 $C_1$，$C_2$ 且满足 $\langle C_1,C_2\rangle\in R_c$。

**定义 4** 相邻概念之间的相似度，表示两个相邻概念之间联系的强度，其计算方法采用

$$\mathrm{Sim}(C_1,C_2)=P(C_1\cap C_2)/P(C_1\cup C_2)$$

其中 $P(C_1\cap C_2)$ 表示在知识库中概念 $C_1$ 和 $C_2$ 同时出现的概率，而 $P(C_1\cup C_2)$ 表示知识库中包含概念 $C_1$ 或者 $C_2$ 的概念。

那么相邻概念之间的距离就可以表示为：

$$\mathrm{dist}(C_1,C_2)=\begin{cases}1/\mathrm{Sim}(C_1,C_2), & Sim(C_1,C_2)\neq 0\\ \infty, & \mathrm{Sim}(C_1,C_2)=0\end{cases}$$

本体中所有的概念按照相邻关系可构成一个全连通的图，即任意两个概念 $C_1$ 和 $C_n$ 之间存在一条路径 $C_1,C_2,\cdots C_k,\cdots,C_n$。那么 $C_1$ 和 $C_n$ 之间的距离可以使用两个概念之间的最短路径的长度来表示。假设 $C_1,C_2,\cdots C_i,\cdots,C_n$ 是两个概念之间的最短路径，那么 $C_1$ 和 $C_n$ 的距离就是：

$$\mathrm{dist}(C_1,C_n)=\sum_{i=1}^{n-1}\mathrm{dist}(C_i,C_{i+1})$$

**定义 5** 对于任意两个非相邻概念 $C_1$，$C_n$，它们之间的最短路径为 $C_1,C_2,\cdots C_k,\cdots,C_n$，那么两个概念之间的相似度可以表示为：

$$\mathrm{Sim}(C_1,C_n)=\prod_{i=1}^{n-1}\mathrm{Sim}(C_i,C_{i+1})$$

此外，本文对同一个概念之间的距离和相似度分别定义为 0 和 1。

## 4 基于本体的文本分类算法

文本分类系统的任务就是在给定的分类体系下，根据文本的内容或属性，将大量的文本归到一个或多个类别中。从数学角度来看，文本分类是一个映射的过程，它将未标明类别的文本映射到已有的类别中。对于知识管理系统，本体集中的概念本身就是知识管理中的一种分类体系，面向知识管理的文本分类过程就是判断待分类文本涉及和描述了本体集中哪些概念，并将该文档划分到概念所对应的类别中。

基于本体的文本分类算法描述：

输入：本体集 Ontology，Ontology 中任意两个概念之间的相似度矩阵，以及待分类文本 W_T。

输出：W_T 所属类的隶属度向量 V。

Step 1：文本的预处理和初始化操作。使用分词算法将待分类的文本文件 W_T 转换为词向量的形式 W_V=$\{w_0,w_2,\cdots,w_n\}$；令 V=0，初始化用来统计每个概念在文本中出现次数的向量 $V_{\mathrm{count}}=0$，以及属于每个概念的强度向量 $V_s=0$；

Step 2：统计每个概念在文档中出现的次数，如果在文档中所有的概念都没有出现，那么只跳转的 Step6；

Step 3：计算该文本隶属于每个概念的强度：

$V_s(i)=\sum_{j=0}^{m}V_{\mathrm{count}}(i)\times\mathrm{Sim}(C_i,C_j)$

Step 4：计算该文本对于每个概念的隶属度：

$V(i)=V_s(i)/\sum_{j=0}^{m}V_s(j)$

Step 5：根据概念之间的语义关系调整隶属度向量：如果 $V(i)>V(j)$ 且概念 $C_i$ 和 $C_j$ 之间存在 kind-of，instance-of，part-of，attribute-of 四种明确的语义关系，那么令 $V(i)=V(j)$。

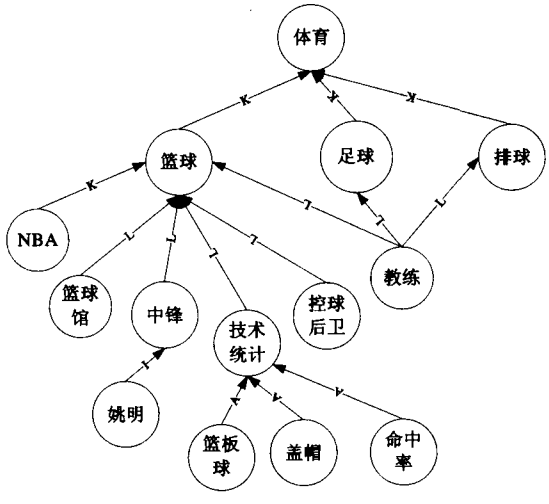
Step 6：输出结果向量。

## 5 实验及结果分析

### 5.1 实验数据集及方法

为了验证本文提出方法的可行性和有效性，本文使用搜狐实验室提供的已分类的语料库。本文使用了其中教育、财经、健康、体育、军事等五类文档，每类文档 1990 个，共计 9950 个文档。实验过程中利用每一个文件中的 1000 个来训练和计算概念之间的相似度，并在另外的 990 个文档中进行测试。实验过程中的分词技术采用 2-gram 方法，训练数据选自北京大学计算语言学研究所提供的人民日报标注语料库[8]。为了进一步提高分词精度，实验过程中又添加了 50 篇语料，及每一类文本选择了 10 篇语料。

由于新闻网站的每类文档涉及面较广，因此难以构造完备的本体集。本文在实验过程中通过统计方法，在各类文档数据中各筛选 300 个高频的名词和动词作为候选概念，最后分别为教育、财经、健康、体育、军事中各选 100 个概念，并建立概念之间的关系。图 1 所示的是一个体育类本体中概念之间存在的部分关系。




（K：表示 kind-of 关系，I 表示 instance-of 关系，A 表示 Attribute-of 关系，L 表示 Link-of 关系）

图 1 描述体育的本体集

### 5.2 实验结果及其分析

对于文本分类算法的评估主要考察算法的查全率 $P_{recall}$ 和查准率 $P_{precision}$ 两个指标。查全率是指正确分类到某类文档的数量 $K_n$ 与该类所有文档的数量之比 $K_a$。查准率是指正确分类到该类文档的数量与所有划分到该类的文档数量 $K_p$ 之比。

$$P_{recall}=\frac{\text{正确分类的文本数量 } K_n}{\text{该类文本总数 } K_a}$$

$$P_{precision}=\frac{\text{正确分类的文本数量 } K_n}{\text{所有划分到该类文本的数量 } K_p}$$

为了比较算法分类的性能，实验同时采用了 KNN 分类方法作为比较，KNN 方法采用词向量空间模型来表示文本文件的特征，每一类文本选择了 1000 维特征向量。两种文本分类方法的分类的结果见表 2 和表 3。

表 2 基于 KNN 的分类混淆矩阵

| 数量 | 教育 | 财经 | 健康 | 体育 | 军事 |
|---|---|---|---|---|---|
| 教育 | 734 | 109 | 58 | 17 | 1 |
| 财经 | 112 | 751 | 200 | 38 | 69 |
| 健康 | 37 | 63 | 690 | 165 | 30 |
| 体育 | 74 | 49 | 45 | 765 | 46 |
| 军事 | 43 | 28 | 7 | 15 | 854 |
| 查准率 | 0.798 | 0.641 | 0.700 | 0.781 | 0.901 |
| 查全率 | 0.734 | 0.751 | 0.69 | 0.765 | 0.854 |

表 3 基于本体的分类混淆矩阵

| 数量 | 教育 | 财经 | 健康 | 体育 | 军事 |
|---|---|---|---|---|---|
| 教育 | 854 | 33 | 58 | 22 | 19 |
| 财经 | 63 | 871 | 84 | 38 | 34 |
| 健康 | 29 | 26 | 763 | 65 | 31 |
| 体育 | 19 | 49 | 55 | 823 | 44 |
| 军事 | 35 | 21 | 40 | 52 | 872 |
| 查准率 | 0.866 | 0.799 | 0.834 | 0.831 | 0.854 |
| 查全率 | 0.854 | 0.871 | 0.763 | 0.823 | 0.872 |

从表 2 和表 3 中的文本分类结果可以看到，本文提出的文本分类方法与 KNN 分类方法在大多数类别上的查准率和查全率都有了一定程度的提高，而且分类性能波动性较小。

特别是基于本体的分类方法，能够将文档定位到更为精确的类别上，例如新的分类器不仅可以识别某文档是体育类别的文档，而且可以区分是属于篮球方面的还是足球方面的文档。因此实现多粒度的分类有助于知识管理系统存储和检索企业内的各种知识。

新的文本分类方法思想简单，可以将文本分类过程转化为受控关键词的统计问题。因此算法不需要进行特征选取和高维特征向量的计算。

尽管实验过程中构建的本体还很粗略，只选取了部分频率较高的概念，而且本体集中包含的概念也不完整，但是已经具备了良好的分类性能，而且其分类结果包含了较多的语义信息。

**结论** 本文研究了一种新的面向知识管理的文本分类方法，该方法从知识管理具备的特点，设计了一种基于本体的文本分类技术。实验结果表明新方法具有良好的分类性能，同时分类的结果具有更为丰富的语义信息。

## 参 考 文 献


1 Fischer G, Otswald J. Knowledge management: problems, promises, realities, and challenges [J]. IEEE Intelligent Systems and Their Applications, 2001,16(1):60～72

2 Kim S B. Some Effective Techniques for Naive Bayes Text Classification [J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(11):1457～1466

3 Vapnik V. The Nature of Statistical Learning Theory [M]. New York: Springer, 2000

4 Bell D A, Guan J W, Bi Y. On combining classifier mass functions for text categorization. IEEE Transactions on Knowledge and Data Engineering, 2005, 8(10): 1307～1319

5 Lam W, Han Yiqiu. Automatic textual document categorization based on generalized instance sets and a metamodel [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003,25(5):628～633

6 刘柏嵩. 基于本体的知识管理关键技术研究[J]. 情报学报, 2005,24(1): 75～81

7 Staab S, Studer R. Knowledge processes and ontologies [J]. IEEE Intelligent Systems, 2001, 16(1): 26～34

8 http://www.icl.pku.edu.cn/icl_res/