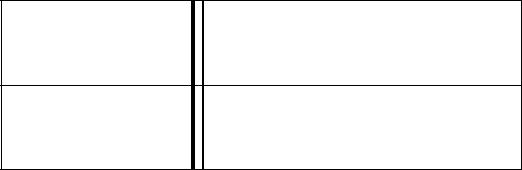
《智能信息处理》课程考试

**本体在文本自动分类中的应用**

|  |  |  |
| --- | --- | --- |
| **学 院：** | 信息科学技术学院 | |
| **专 业：** | 计算机科学与技术 | |
| **姓 名：** | 王云涛 | |
| **学 号：** | 1120200304 |  |
| **授 课 老 师：** | 李冠宇 | |



考核 课程成绩

得分

**本体在文本自动分类中的应用**

**王云涛**

（大连海事大学信息科学技术学院 大连 116026）

**摘 要** 随着计算机技术的飞速发展，各领域文本信息呈现爆炸式增长，这使得文本自动分类技术越来越得到青睐。对比传统文本自动分类技术，基于本体的文本自动分类将本体与文本自动分类技术结合，并在此基础上提出了 OOMD 模型。基于本体的构建方法使得文本的语义信息易于掌握，分类的准确度也相应提高。

**关键词** 本体，文本自动分类，OOMD 模型中**图法分类号** TP393 **文献标识码** A

# Ontology in Automatic Text Classification

Wang Yuntao

(Information Science and Technology, Dalian Maritime University, Dalian 116026, China)

**Abstract** With the rapid development of computer technology ,various fields of text information showing explosive growth, which makes automatic text classification technology more and more popular. Compared to traditional automatic text classification technology, ontology-based automatic text classification combines ontology and the automatic text classification technologies, on which the model OOMD was proposed. Ontology-based construction method makes semantic information of the text easily to grasp, the classification accuracy is also increased accordingly.

**Keywords** Ontology, Automatic Text Classification,OOMD Model

# 引言

自 1946 年第一台计算机诞生以来，计算机技术和信息技术以惊人的速度向前发展，进而衍生了互联网技术的应用。1990 年，Tim Berners-Lee 发明了万维网，自此人们可以通过因特网（Internet）来获得各种信息。互联网技术的应用使得各式各样的信息层出不穷，人们已经逐渐从信息匮乏的时代走入信息过载的时代。信息以网络为载体，存储为文本、图像、音频、视频等。就文本而言，每天各领域产生的文本信息复杂多样。依据 DBLP 的统计数据，自 1970 至 2010 年这40 年间，每年学者发表的论文数量呈现指数增长，仅 2010年一年，各领域学者发表的论文竟有 16000 篇之多。各个领域的文本信息复杂多样并且数量庞大，而对文本信息的处理 又是挖掘客观规律，做出利于企业发展的决策的必经之路。由此可见，如何在各类文本信息中寻找满足需求的信息显得尤为重要，这就涉及到对文本进行分类。搜索引擎很难满足不同用户对检索结果精化的要求[4]。分门别类的文本信息更有利于检 索的实现，也更满足网民的需求。

对文本进行有效的管理方法之一就是将它们进行系统的分类[3]。文本自动分类（Automatic Text Classification）的

应用使得文本信息的类别更加清晰化，更对网络上纷繁复杂 的大量文本信息有了一个基本的分类，使得在检索过程中对 于目标文档的查找更为容易。目前已有成型的文本自动分类 算法的出现，这些算法大部分是有监督学习的机器学习算法。或者给定两类或者给定多类，对样本进行训练后进行分类。文本自动分类技术解决了信息过载的问题，同时也推进了诸 如雅虎这类以分类目录为核心的推荐系统的产生和发展。

传统的自动文本分类技术大多针对分类的准确率和查全率，基本上是相似性的匹配算法，并不能满足用户在语义和知识上的需求。本体具有良好的层次概念结构和对逻辑推理的支持，本文采用本体技术构建自动文本分类，很好的解决了分类的语义上的弊端。基于本体的文本自动分类技术的应用，在一定程度上保证了准确率和查全率，同时也对文本的语义进行涵盖，更加适应社会的发展趋势。

# 本体

## 本体的定义

本体（Ontology）原来是一个哲学观念，本意是指事物的本质。哲学上一般称本体为本体论，是在 17 世纪由西方哲学家提出的，表示存在及其本质和规律。在以后的发展中， 本体逐渐成为研究实体存在性和实体存在的本质等方面的

. 1.

通用理论。工程学学者将其用来在开发知识系统时用于领域知识的获取，人工智能领域则用其给出构成相关领域词汇的基本术语和关系。

1993 年，Gruber 给出本体的定义：“本体是概念模型的明确的规范说明。”1997 年，Borst 对此进行了修改，提出：“本体是共享概念模型的形式化规范说明。”1998 年，Studer 等对上述两个概念进行了深入的研究，并对其进行了扩充， 认为本体是共享概念模型的明确的形式化规范说明。这个定义包含 4 层含义：概念化、明确化、形式化和共享。

概念化是指通过抽象出客观世界中的一些现象的相关概念得到的蕴含语义结构的概念模型，表达为一组概念、定义和关系。明确化指使用的概念以及使用这些概念的约束都有明确的定义。形式化指本体能够被计算机所处理。共享指本体针对社会范畴而非个体之间的共识。

一般来讲，本体有两个特性：静态性和动态性。静态性是指本体反映的是概念模型，并没有涉及动态的行为。动态性指的是本体服务的对象和内容是不断变化的，针对不同领域有不同的本体可以构造出来。

## 本体的分类

本体的类型分为以下四种。

* + 1. 通用知识本体：从概念的根结点出发来进行抽象。例如 ONTOL-MT 这个通用知识本体的初始概念有事物、时间、空间、数量、行为状态和属性共 6 个下位概念。
    2. 领域知识本体：对领域知识的抽象，概念明确， 容易形式化和共享。
    3. 语言知识本体：一个描述单词和术语之间的概念关系的词表。
    4. 形式知识本体：对概念和术语的分类很严格，要按一定的原则和标准，明确地定义概念之间的显性和隐性关系，明确概念的约束和逻辑联系。

# 文本自动分类方法

自动分类就是用计算机代替人工对文献等对象进行分类[1]。分类算法是一种有监督的学习算法[2]。首先设立一组已经预先分类的文档作为训练集合，对其进行分析得到分类模型，再使用分类模型对待测试样本进行训练。目前，主要的文本自动分类算法可以分为三类：词匹配法、基于知识工程的方法和统计学习法。

1. 词匹配法：根据文本中与类别中共同出现的词来决定该文本属于哪种类别。该分类方法对语境不敏感，无法正确处理在不同的上下文环境中某个词的不同含义，分类的准确度也很低。
2. 基于知识工程的方法：该方法依赖于知识工程师 手工编写大量的的推理规则，这些规则通常面向具体的领域， 当处于不同领域的分类问题时,需要不同领域的知识工程师

编写不同的推理规则，并且分类质量严重依赖于推理规则的质量，在实际的分类系统中较少使用基于知识工程的学习法。

1. 统计学习法：基本思路是先搜集一些与待分类的文本属于同一领域的文档作为训练集，并由专家进行分类， 保证分类的准确性，然后分析这些已经分好类的文档，从中挖掘关键词和类别之间的联系，最后再利用这些学到的知识进行分类，而不是机械地按词进行匹配。这种基于统计的经验学习法具有良好的理论基础，简单的实现机制以及较好的分类质量，目前实用的分类系统基本上都是采用这种分类方法。

根据分类的结果不同,基于统计学习法的分类系统又分为：独立二元(Independent Binary)分类系统和 *m* 元(m-ary) 分类系统。独立二元分类就是给定一篇文档分类系统对每一个分类都进行判断这篇文档是否属于这个类别，其结果只有两个，要么属于，要么不属于。 *m* 元分类系统就是给定一篇文档，计算该文档与每个类别的相似度,最后按照相似度对类别进行排序。国外当前流行的文本分类方法有 Rocchio 法及其变异方法、k 近邻法(kNN)、决策树、朴素贝叶斯、贝叶斯网络、支持向量机(SVM)等方法。对于文本自动分类方法的汇总如图 1。

文本自动分类算法

词匹配法

统计学习法

知识工程法

Independency Binary

M-ary

DTree NB NN NNet Rocchio SVM LLSF

KNN

WORD

图 1 文本自动分类算法汇总图

.2.

# 基于本体的文本自动分类系统

## 文本集的本体表示

从本质上说，本体是概念的形式化、显示规范，是结构化的概念集。本体可以表示为偏序集合<*T*,≤,*R*>。其中，*T* 为术语集，≤为概念之间的偏序关系，*R* 为概念之间的公理化约束。本体也可以用图模型<*N*,*E*,*R*>表示。概念集 *N* 构成图中的节点集合，概念之间的关系构成图中的边的集合 *E*，*R* 构成概念之间的公理约束。本体通过对概念的严格定义和概念之间的关系来确定概念的精确含义，表示共同认可的、可共享的知识[5]。

本 文 设 计 了 面 向 本 体 的 文 本 模 型

（OOMD:Ontology-Oriented Model of Documents），如图 2 所示，该模型采用本体作为文本内容的描述方法，为后续的分类器提供一个涵盖语义特征的概念模型，进而提供一种在语义和知识层次上描述文本进而进行分类的方法。

念，每个概念被描述成一个具有文档固有信息以及语义信息的概念本体，作为 OOMD 模型的原子单元进行训练。

（2）计算概念本体之间的相似性度量。对涵盖语义的概念本体进行相似性度量，初步判断各文档的类别，进行领域分类。领域本体是对领域概念化的一个详细说明[6]。在处理某一个概念本体时，使用余弦相似性度量来衡量概念本体之间的相似度。

## 基于本体的文本自动分类

本文通过对本体的解析和应用在语义和知识层面上实现了文档归类，并以此为依据设计了文本自动分类系统。该系统针对各个本体概念选择训练样本，使用分类器来对各个概念本体进行训练，以实现文档的归类。系统分为文档预处理和分类器训练两个模块，其结构流程图如图 3 所示。

训练样本库

原始文档库

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

|  |  |
| --- | --- |
|  |  |
|  |  |

分词 去停词



本体集成

规则库

关系识别

语义解释

**OOMD**

知识库

概念形成

领域分类

主题词提取

文档集合

提取主题词

统计词频

分词 去停词

提取主题词

统计词频

**文档预处理**

概念本体

领域本体

关系识别

图 2 面向本体的文本模型 OOMD

OOMD 的基本思想是：首先对文档集合进行主题词的提取，然后将文档中的主题词按照语义进行领域分类，并利用本体建模的方法建立不同分类的概念本体，解释概念蕴含的语义，识别概念本体之间的关系，形成规则库，最后将概念本体及其关系集成为一个表示一定语义信息的领域本体，其主要步骤如下：

（1）提取文档主题词。在文档集合中提取主题词，把

知识库 规则库

分类器训练

用户

**分类器训练**

文档集中的 *n* 个主题词表示为 *x1*，x2，…，*xn*，每个文档的主题词表示为一个整体 *X*=[*x1*，*x2*，…，*xn*]，作为一篇文档的标识。*X* 是提取主题词过程中最基本的原子单元，代表一个概

图 3 文本自动分类系统的结构流程

文档预处理模块对原始文档集合及训练样本进行了预处理，包括进行分词，英文文档的词性标注，去除停用词， 进行词频统计等。利用隐含狄利克雷分布（LDA：Latent

.3.

Dirichlet Allocation）算法，对文档集合进行主题词提取。分类器训练模块则构造整个文档集的主题词的向量模型，对每一篇文档中的主题词组成的向量作为一个概念本体， 某一类文档集的概念本体的全部作为一个领域本体。对比训练样本中的主题词模型和本体中的概念，设定主题词的权重， 即一个词预测主题的能力越强，其权重就越大，反之，权重就越小。在使用分类器训练样本时，计算原始样本概念本体 及训练样本之间的概念本体之间的相似度，对训练样本进行分类。

本文通过对文本自动分类的模型的描述、文本预处理方法以及其分类器训练进行了研究，采用领域本体作为文本自动分类的概念模型，提出来一种基于本体的文本自动分类的方法，实现了从语义和知识层次上对文本进行归档，为文本信息处理提供了一种新的实施方案。

**结束语** 对于文本自动分类的研究，现今已有很多成型的研究成果，其中基于统计学的方法广为应用，在这个领域，已经有比较成熟的方法在实际应用。本文从概念方面进行入手，采用本体论的方法，在文档集的概念模型层次上，利用文档集的语义特征和知识层次来组织文档的主题词，对比传统基于统计学的文本自动分类方法，更智能化得表达了文档的语义特征，为文档自动分类方法提供了一个新颖的思路。

# 参 考 文 献

1. 王涛 . 文 本 自 动 分 类 研 究 [J]. 图 书 馆 学 研究,2007,12(9):40-46.
2. 盛秋艳. 文本自动分类技术的研究[J]. 交通科技与经济,2006 ,3 (7 ) : 92 - 124 .
3. 王宏生, 张琳. 基于本体的文本自动分类[J]. 科技信息,2008,29(8):549-552.
4. 袁晓曦.基于机器学习的 Web 文本自动分类[J].软件导刊,2011,10(1):26-28.
5. 郭润寰.基于本体的语义网技术研究与应用[D].陕西:西安电子科技大学,2005.
6. 宋峻峰,张维明等.基于本体的信息检索研究[J].南京大学学报,2005,41(2):189-197.

.4.