Research of Text Categorization Based on Ontology

WANG Jiayun, ZHANG Rui, WANG Peng
Chengdu University of Information Technology, Chengdu, China
e-mail: *wjy@cuit.edu.cn*

Abstract: In order to get optimal text categorization, semantic information must be saved mostly. Based on domain ontology, an efficiency concept text categorization was proposed and implemented. Using ontology hierarchy and attribute constraint, keywords matched domain ontology, which builded concept vector space model for text categorization. The advantage of this method resolved polysemy, synonym, and concept hierarchy problem.

Keywords: ontology; concept vector space model; text categorization

基于Ontology的概念向量空间模型文本分类研究

王嘉昀，张睿，王鹏
成都信息工程学院软件工程学院，成都，中国，610225
e-mail: wjy@cuit.edu.cn

【摘　要】文本分类要取得理想分类效果，必须极大限度保留文本的语义信息。本文将领域本体引入到文本分类中，构造了基于概念的文本分类系统架构。利用本体类层次结构及属性约束等特点，将关键词与领域本体概念进行匹配，建立概念向量空间模型进行文本分类，这有利于解决文本分类中术语一词多义、一义多词和概念的层次问题。

【关键词】Ontology；概念向量空间模型；文本分类

中图法分类号: TP391　　文献标识码: A

1 引言

传统文本分类通常采用关键词加权重的向量空间模型对文本进行分类，其优势在于简便快捷，但当系统的性能达到一定程度后，无论怎样改进分类算法，性能都很难再得到提高。究其原因主要是由于现实文本中的用词往往都是有语义关联的，如同义关系、近义关系、上下位关系等。为克服传统分类方法中基于关键词匹配的局限性，本文尝试引入领域本体，借助本体类层次结构及属性约束等特点，提出基于Ontology 的概念向量模型文本分类方法。

2 Ontology 概述

2.1 本体定义

本体(Ontology)是一个哲学概念，是对世界上客观存在物的系统地描述，关心的是客观现实的抽象本质。在人工智能界，最早给出Ontology定义的是Neches等人，他们将Ontology定义为“给出构成相关领域词汇的基本术语和关系，以及利用这些术语和关系构成的规定这些词汇外延的规则的定义”。1993年，Gruber给出了Ontology的一个最为流行的定义[1]，即“Ontology是概念模型的明确的规范说明”。后来Borst对这个定义进行引申，提出“Ontology是共享概念模型的形式化规范说明”[2]。Studer等人对上述两个定义进行了深入的研究，认为Ontology是共享概念模型的明确的形式化规范说明。他们认为本体概念包括4个主要方面[3]：概念模型(Conceptualization)、明确(Explicit)、形式化(Formal)和共享(Share)。

2.2 本体语言

本体可以采用多种形式来表示，但是一般都包含一个领域的词汇表和词汇意义的某些说明。在最简单的情况下，本体可以只描述由包含关系关联起来的概念层次。

近几年来，许多研究者已经开发出多种本体语言，用于实现在异构环境中异构数据的交换。比较有代表性的有CycL、Ontolingua、XOL、SHOE、RDF(S)、OIL、DAML+OIL、OWL等。其中，

978-1-935068-06-8

DAML+OIL语言，采用面向对象的方法，用类和属性来描述领域的结构，具有清晰的语义，成为W3C研究语义Web中Ontology描述语言的起点，并成为标准。本文将应用该语言对本体中可实现的各种语义关系进行说明。

2.3 本体中语义关系的实现

人们看待事物的角度不同会导致对同一概念的不同表达形式，即使用不同的词汇表达相同或相近的主题概念，或者是为了避免使用同样的词汇而使用同义词替换，因此在对特征词或主题词进行匹配的过程中应该考虑到应用词语间的语义关系对特征词进行扩展。在主题分类时可以利用词与词之间存在的3种语义关系[4,6]：

（1）同义关系（synonym）：词与词之间的意义相同或非常相近，往往可以互相替代。如：计算机和微机的关系。DAML+OIL中可以有如下实现方法：

< daml:Class rdf : ID="计算机"/>
< daml:Class rdf : ID="微机 ">
< rdfs:sameClassAs rdf : resource ="计算机">
< / daml: Class >

（2）上下位关系（hypernym/hyponym）：又被称为“is-a”，下位词是上位词的特例，如设备和计算机、投影仪、录音机之间的关系。在DAML+OIL中可以表示为：

< daml:Class rdf : ID="设备"/>
< daml:Class rdf : ID="计算机">
< daml:Class rdf : ID="投影仪">
< rdfs:subClassof rdf : resource ="设备">
< / daml: Class >

此外，本体中概念的层次结构也是由这种上下位关系体现的。

（3）包含关系（meronym/holonym）：又被称为“has-a”，即部分与整体的关系，如计算机由主板、CPU、内存、硬盘等组成的。

<daml:Class rdf:ID="计算机">
<rdf:subClassOf>
<daml:Restriction daml:cardinality=1>
<daml:onProperty rdf:resource="hasKeyBoard"/>
<daml:toClass rdf:resource="主板" />
</daml:Restriciton>
</rdf:subClassOf>
<rdf:subClassOf>
<daml:Restriction>
<daml:onProperty rdf:resource="hasMemory"/>
<daml:toClass rdf:resource="内存" />
</daml:Restriciton>
</rdf:subClassOf>
<rdf:subClassOf>
<daml:Restriction >
<daml:onProperty rdf:resource=" hasHard" />
<daml:toClass rdf:resource="硬盘"/>
</daml:Restriciton>
</rdf:subClassOf>
</daml:Class>
<daml:ObjectProperty rdf:ID= "hasKeyBoard" />
<daml:ObjectProperty rdf:ID= "hasMemory" />
<daml:ObjectProperty rdf:ID= "hasHard" />

3 基于概念向量模型的文本分类

3.1 文本分类系统框架

文本分类系统包括预处理、匹配模块、特征提取、特征表示、分类算法及结果评测。基于概念向量文本分类系统架构如图1所示。其中匹配模块引入领域本体，将关键词与领域本体中概念进行转化。

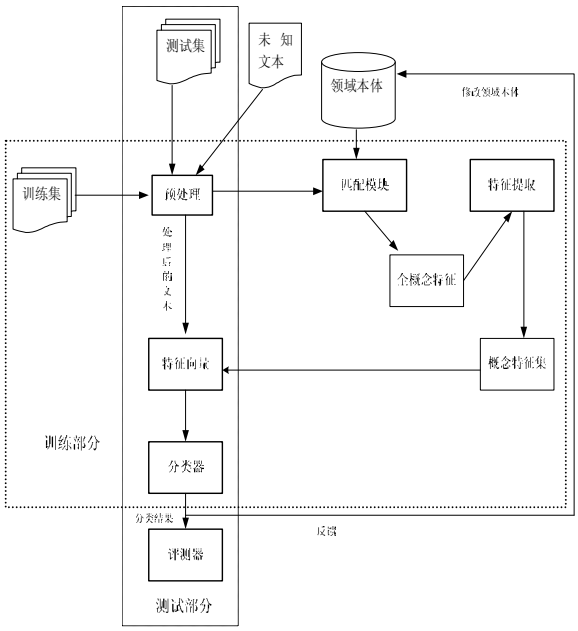

Figure 1. The system frame of text categorization

图1.文本分类系统框架

3.2 预处理

预处理模块主要完成对文本分词和词性标注，并通过禁用词表除去出现频率高但对文本分类不起作用的词。本文直接采用中科院计算所汉语词法分析系统 ICTCLAS(Institute of Computing Technology, Chinese Lexical Analysis System)。

3.3 匹配模块

匹配模块是利用如下匹配算法，将预处理得到的关键词与领域本体中的概念进行匹配，匹配成功，获取本体中的概念，用本体中概念代替该关键词；匹配不成功，表示该关键词不满足本体中任何概念，对于这样的关键词我们将其作为未登录词处理。具体步骤如下：

(1)将分词后的文本 d={k_1 ,k_2 ,Λ,k_n}中的第 i 个关键词 k_i 送入到领域本体中，将其与本体中的属性进行匹配。因为本体中的概念是通过属性集合进行定义的，而这些属性实际上就是一些关键词，所以通过匹配属性来获取概念。形式化表示为：

$$f_{(d,c_j)}(k_i, a_{jt})=\begin{cases}1,\cdots if\ k_i = a_{jt}\\ 0,\cdots if\ k_i \neq a_{jt}\end{cases}$$

(2)匹配成功，即 $f_{(d,c_j)}(k_i, a_{jt})=1$，表明本体中某个概念 c_j 的第 t 个属性与 k_i 相同，则我们用概念 c_j 来代替关键词 k_i 。

(3)若不能匹配，即 $f_{(d,c_j)}(k_i, a_{jt})=0$，表明关键词 k_i 不能满足本体中的任何概念，对于这样的关键词直接保留，将其作为未登录词来对待。

通过上述方法，我们可以完成关键词与本体匹配，得到本体中的概念，从而将文本表示成概念的集合。

3.4 概念特征提取

特征提取就是对全部概念特征及未登录关键词进行冗余处理的过程。提取后的概念特征将用于之后的训练和分类过程。概念特征提取的形式化描述为：

(1)如果存在概念 $c_1^{k_1}, c_2^{k_2}, \Lambda, c_i^{ki}$,有 $c_1^{k_1}=c_2^{k_2}=\Lambda=c_i^{ki}$ ，则保留 $c_1^{k_1}$ 而删除 $c_2^{k_2}, \Lambda, c_i^{ki}$ 。

(2)如果存在 $c_i^{ki}=k_i$ ，则表明 k_i 是一个未登录词。然后根据关键词权重计算方法，如TF-IDF算法计算 k_i 的权重。若其权重大于事先设定阈值 λ_d ，则保留；否则将该关键词去除。

3.5 文本概念特征表示

定义 1：概念特征表示采用传统向量空间模型的思想，即用(concept，value)值对把文本表示为向量空间，我们称为概念向量空间模型(Concept Vector Space Model，简记 CVSM)。concept 部分随着概念特征确定得以确定，剩下的工作就是依据 concept 来计算对应的 value 值。概念权重的计算公式为：

$$w(c_j^{k_j})=\lambda^n \times w_k(c_j^{k_j}) \quad (1)$$

其中，λ(0＜λ＜1)为指数因子，n 为概念在领域本体的层次，而对于 $w_k(c_j^{k_j})$ 的计算，我们将分情况确定：

(1)若 $c_j^{k_j}$ 是领域本体中的一个概念，并且是通过文档中的 p 个关键词匹配得到的，那么其权重为 p 个关键词权重之和，即

$$w_k(c_j^{k_j})=\sum_{i=1}^{p} w_k(k_i) \quad (2)$$

(2)若 $c_j^{k_j}$ 是领域本体一个未登录关键词，则概念的权重即为关键词的权重，即

$$w_k(c_j^{k_j})=w_k(k_j) \quad (3)$$

上述 $w_k(k_j)$ 为关键词在文档中的权重，常用关键词权重的计算方法有布尔值算法、词频算法和 TF-IDF 算法，其中 TF-IDF 算法最能反映关键词的权重。

3.6 文本相似度计算

对于两个文本间的相似度计算，我们将分为两个部分进行分析：对于文本的概念向量空间模型，我们将其表示成 CM=C∪S，这里 C 是概念集合而 S 是未登录词的集合。我们设 x 和 y 是文本集 D 中任意两个文本，文本集中文本总数为 N。

定义 2：设 c 和 c' 是领域本体中的两个概念，则这两个概念之间的距离为概念间的语义相似度，其计算公式如下：[5]

$$s(c,c')=\frac{1}{l(c,c')\log(D(c)+D(c')+1)} \quad (4)$$

其中，$l(c,c')$ 为概念 c 和 c' 之间的层次距离，$D(c)$ 和 $D(c')$ 分别为概念 c 和 c' 的子孙节点数目。特别指出，对于所有概念 s(c, c)=1，$l(c,c')=l(c',c)$，$D(c)\geq 0$，$D(c')\geq 0$，并且 $D(c)$ 、$D(c')$ 至少有一个不

978-1-935068-06-8 © 2009 SciRes.

小于 1。

(1) 基于概念向量空间模型文本相似度计算

定义 3：基于概念文本 x 和 y 之间的内积 $<x,y>^c$ 定义如下：

$$<x,y>^c = \sum_{c\in C}\sum_{c'\in C} l_c \tau_{c,x} l_{c'} \tau_{c',y} s(c,c') \qquad (5)$$

其中，l_c 和 $l_{c'}$ 分别为包含概念 c，c' 的文档数的倒数，$\tau_{c,x}$ 为概念 c 在文档 x 中出现的次数，$\tau_{c',y}$ 为概念 c' 在文档 y 中出现的次数，s(c, c')为概念 c 和 c' 的语义相似度。基于概念文本内积，我们就可以定义文本 x 和 y 的概念向量相似度计算公式：

$$\mathrm{Sim}^c(x,y) = \frac{<x,y>^c}{\sqrt{<x,x>^c <y,y>^c}} \qquad (6)$$

(2) 基于关键词向量空间模型文本相似度计算

定义 4：基于关键词文本 x 和 y 之间的内积 $<x,y>^s$ 定义如下：

$$<x,y>^s = \sum_{s\in S} l_s^2 \tau_{s,x} \tau_{s,y} \qquad (7)$$

其中，l_s 为包含关键词 s 的文档数的倒数，$\tau_{s,x}$ 为关键词 s 在文档 x 中出现的次数，$\tau_{s,y}$ 为关键词 s 在文档 y 中出现的次数。基于关键词文本内积，我们就可以定义文本 x 和 y 的文本相似度计算公式：

$$\mathrm{Sim}^s(x,y) = \frac{<x,y>^s}{\sqrt{<x,x>^s <y,y>^s}} \qquad (8)$$

综合上面两部分的分析，我们可以得到基于概念向量的两个文本间的相似度的计算公式，如下：

$$\mathrm{Sim}(x,y) = \alpha \cdot \mathrm{Sim}^c(x,y) + (1-\alpha) \cdot \mathrm{Sim}^s(x,y) \qquad (9)$$

其中，参数 α 的作用是调节概念和未登录词对文本间相似度计算所产生的不同影响。

4 结论

本文将领域本体引入到文本分类中，构造了基于概念的文本分类系统架构。利用本体类层次结构及属性约束等特点，将关键词与领域本体概念进行匹配，建立概念向量空间模型进行文本分类，解决文本分类中术语一词多义、一义多词和概念的层次问题。

Reference (参考文献)

[1] Gruber T R. A Translation Approach to Portable Ontology Specifications.Knowledge Acquisition.1993,5:199~220.

[2] 邓志鸿，唐世渭，张铭，杨冬青，陈捷.Ontology 研究综述[J]. 北京大学学报(自然科学版)，2002，38(5):730~738.

[3] Studer R, Benjamins V R, Fensel D. Knowledge Engineering: Principles and Methods. Data and Knowledge Engineering, 1998, 25(1-2):161~197.

[4] 吴晓，李丹宁.基于综合倒排索引的个性化搜索引擎研究[J] . 微计算机信息，2008,9：201-203.

[5] W.Mao,W.W. Chu. The Phrase-based vector space model for automatic retrieval of free-text medical documents.Data &Knowledge Engineering 61(2007) 76~92.

[6] http://protegewiki.stanford.edu/index.php/Protege_Ontology_Library.