·人工智能及识别技术· 文章编号:1000-3428(2016)12-0181-07 文献标志码:A 中图分类号:TP391.1

基于概念聚类的领域本体图中文文本分类

叶施仁,孙 宁

(常州大学 信息科学与工程学院,江苏 常州 213164)

摘 要: 基于半监督概念聚类技术,提出一种改进的领域本体图中文文本分类算法。根据领域本体图结构模型,创建中文文本分类的本体学习框架,利用 HowNet 字典实现术语提取并建立中文术语-术语关系映射。依据术语间的权重连接关系,设计二分类关系的 KLSeeker 本体中文文本分类算法,并通过基于概念聚类的本体图半监督学习,实现中文文本的精确分类。实验结果表明,与基于非负张量分解的中文文本分类算法相比,该算法具有更高的分类精度。

关键词: 词消歧;半监督;概念聚类;HowNet 字典;二分类关系;领域本体图

中文引用格式: 叶施仁,孙 宁. 基于概念聚类的领域本体图中文文本分类[J]. 计算机工程,2016,42(12):181-187.

英文引用格式: Ye Shiren, Sun Ning. Chinese Text Classification by Domain Ontology Graph Based on Concept Clustering[J]. Computer Engineering,2016,42(12):181-187.

Chinese Text Classification by Domain Ontology Graph Based on Concept Clustering

YE Shiren, SUN Ning

(School of Information Science and Engineering, Changzhou University, Changzhou, Jiangsu 213164, China)

【Abstract】 This paper proposes an improved Chinese text classification algorithm by Domain Ontology Graph(DOG) based on semi-supervised concept clustering. According to the DOG structure model, the ontology learning framework of Chinese classification is created, and then the HowNet dictionary is used to extract the word disambiguation, and then the Chinese term-term relationship mapping is established. Based on the weight connection between the terms of the binary classification relationship, here designs the KLSeeker ontology Chinese text classification algorithm. It realizes accurate classification of Chinese text through DOG semi-supervised learning. Experimental results show that the proposed algorithm has a higher classification accuracy compared with the Chinese text classification algorithm based on non-negative tensor decomposition ontology concepts.

【Key words】 word disambiguation; semi-supervised; concept clustering; HowNet dictionary; binary classification relationship; Domain Ontology Graph(DOG)

DOI: 10.3969/j.issn.1000-3428.2016.12.032

0 概述

本体图(Ontology Graph,OG)是现实中知识的一种基本表示形式。从计算机科学和信息科学角度来看,本体定义了一组可以建模的具有代表性的原语与知识或话语的域[1]。本体是知识表示语言发展中知识信息系统构建的关键领域[2]。当前有许多代表性的本体学习方法[3-6],可获得所需的本体。然而,上述文献在本体建设和学习过程中,通常需要繁琐的人为干预。这些干预措施可能发生在早期阶段,即运用层次分析法对匹配策略进行选择之前,或是发生在领域本体编辑结束阶段,即纠正或重新使用学习概念过程中。与此同时,一些学者也认为人为干预也可放在算法的中间阶段,构建概念或属性的动态迭代学习方法[7]。由于人为对知识本体的手动创建或维护是耗时和低效的,因此简化的本体学习方法采用最小人为干预,该方法是语义网处理和应用领域较实用和可行的研究方向。

此外,由于文本数据是人类知识的重要来源,利用文本数据进行本体学习是非常有意义的研究工

基金项目: 国家自然科学基金(61272367)。

作者简介: 叶施仁(1970—),男,高级工程师、博士,主研方向为数据挖掘;孙 宁,硕士研究生。

收稿日期: 2015-11-16 **修回日期:** 2015-12-28 **E-mail:** 1113076228@qq.com

作。近年来,许多基于文本数据的本体学习方法得到了广泛的发展。很多研究人员使用机器学习和统计分析方法来构建人工智能本体,并试图从文本数据中自动提取领域本体特征,如文献[8-10]中的本体学习技术对于英语文本数据的学习是比较方便和有效的。然而,由于语言特征不同,应用于英语文本数据的算法在中文文本中的计算效果并不好,汉字结构与英语单词相比更复杂和多元,而且当前尚无成功的中文文本分类系统得到实际应用,因此构建有效的中文文本数据本体学习系统具有理论与实用价值。

从尽可能减少人为参与本体构建的角度,本文开发基于概念聚类的领域本体图中文文本分类算法。针对中文 Web 文本应用领域,该算法可自动生成知识概念化的领域本体图(Domain Ontology Graph,DOG)模型。同时,定义文档本体图(Document Ontology Graph,DocOG)代表单一文本文件内容。DocOG 通过单文档本体与领域本体相匹配,实现基于本体的文本分类。

1 问题描述

本体学习源于哲学学科,已成为计算机科学和信息系统领域的一个热门研究课题。从信息科学的角度来看,本体定义了一组具有代表性的原语,可模拟特定领域或话语的知识。五元组结构是一种常用的概念和域关系的形式化描述方法。五元组本体的核心结构定义为[11]:

$$S=(C,R,H,rel,A) \tag{1}$$

其中,C 为描述对象概念集;R 为关系类型集;H 为对象概念集 C 的关系类型集;$rel\in C\times C$ 为一组与关系型 R 有关的 C 关系集;A 为一组逻辑描述语句。术语 rel 被定义为一组三元组关系:

$$rel=(s,r,o) \tag{2}$$

其中,rel 表示主客体关系;s 为 C 中的主体元素;r 为 R 中的关系元素;o 为 C 中的对象元素。在五元组本体论结构中,知识主要是逻辑语句,并且 rel 为最重要的组成部分,其定义了基于相关概念的关系元组,如图 1 所示。

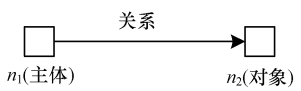


图 1 主体和对象关系

图 1 所示为三元组结构,其主体为源节点 n_1,客体定义为 n_2 靶节点,关系定义为 n_1 和 n_2 之间的关联链接。

文本数据是人类知识最直接的资源,其为描述性数据,使人们能够分享和交换知识。虽然通过计算机对文本数据进行分析并不简单,但近年来许多本体学习方法已得到广泛研究。其中多数使用人工智能方法与半自动文本学习方式,利用信息检索、机器学习、自然语言处理、统计数学等,建立本体学习系统。然而本体学习有时并不能较好地代表人类知识,特别是需要较多的人为干预,这是本文算法需要解决的重点[12]。

2 基于概念聚类的领域本体图分类模型

2.1 本体学习框架

本节中将描述如何生成领域本体图并实现基于文档本体图的文本分类。本体学习框架(KLSeeker)是一个完整的系统框架,定义并实现了4 个组件,主要包括:1)本体图建模;2)本体学习;3)本体生成;4)本体查询。KLSeeker 系统可用来开发各种采用四定义的本体论为基础的智能应用程序组件。因此,整个 KLSeeker 系统框架可分成 4 个模块处理不同种类的本体论,如图 2 所示。

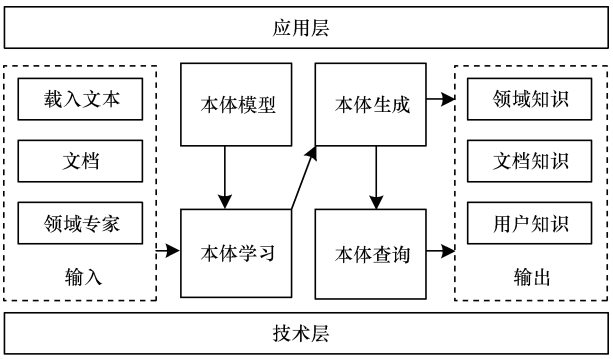


图 2 KLSeeker 系统基本框架

本体图是用于 KLSeeker 系统模型中进行领域知识本体创建的新方法。本体图由不同层次的概念单元组成,由不同类型的关系进行关联,本质是一词汇系统,通过彼此相连表示概念集合。不同概念单元形成网络模型。图 3 进一步给出基于节点和关系结构的本体图模型的概念视图及算法创建和学习过程。

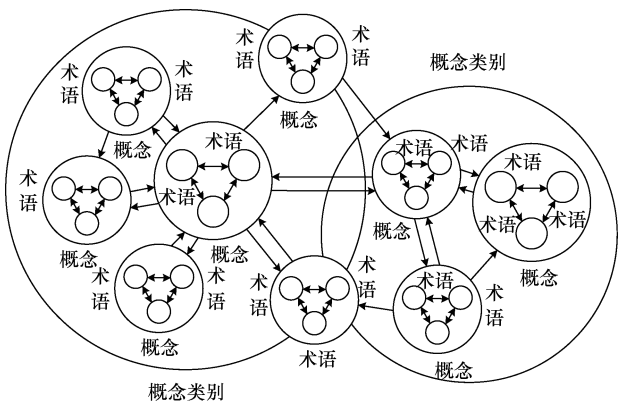


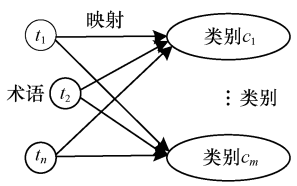
图 3 概念视图

本体图根据知识表现的复杂程度,可分为 4 种类型的概念单元(Concept Unit,CU)。定义任何对象的四客户本体图并给出其语义表达。所有客户通过相关概念进行联系,包括本体图的整个概念结构和模型的区域(领域)知识表达。

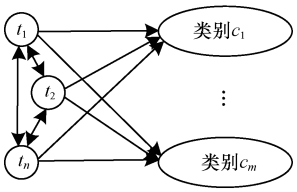
2.2　中文文本本体学习方法

本文主要研究中文文本领域本体图学习算法。由于中文词间无空格,因此英语中可用的词消歧方法在汉语中不可用。中文术语提取通常依赖于字典,HowNet 是目前可用的电子词典,包含超过 50 000 个中文单词,在文本中可构成初始项列表术语提取过程。

通过对单词列表和语料库的中文文本应用最大匹配算法,可提取候选术语列表,同时过滤其他不必要条款。假定提取 n 个候选术语 $T=\{t_1,t_2,\cdots,t_n\}$,其中每个 $t_i(i=1,2,\cdots,n)$ 在文本中至少出现 1 次。从中文文本语料库中提取的候选项列表 T 与本体图模型概念单位没有任何有意义的关联。因此,应用于候选术语表的过程是术语到类的关系映射,如图 4 所示。此映射过程行为,如特征选择,从其最相关的域类术语列表中进行选择和分离。首先,需准备一组标记文本;然后,可衡量条款与类别的关联;最后,从候选词列表中选择子列表,此映射过程意味着可用词和类间加权进行关系表示,如图 4(a)所示。



(a)术语-类别映射



(b)术语-术语映射

图 4　术语关系映射

计算术语 t 和类别 c 间的共现频率 $O_{i,j}$,$i\in\{t,\neg t\}$,$j\in\{c,\neg c\}$,并将其与期望频率 $E_{i,j}$ 进行对比:

$$E_{i,j}=\frac{\sum_{a\in\{t,\neg t\}}O_{a,j}\sum_{b\in\{c,\neg c\}}O_{i,b}}{N} \tag{3}$$

其中,$N=O_{t,c}+O_{t,\neg c}+O_{\neg t,c}+O_{\neg t,\neg c}$ 表示分类文档库大小。术语 t 和类别 c 的独立测量卡方统计定义为:

$$\chi^2_{t,c}=\sum_{i\in\{t,\neg t\}}\sum_{j\in\{c,\neg c\}}\frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}} \tag{4}$$

基于卡方统计的词类依赖测度定义为:

$$R_{t,c}=\frac{O_{t,c}}{E_{t,c}} \tag{5}$$

当 $R_{t,c}>1$ 时,术语 t 为类 c 的正依赖测度;当 $R_{t,c}<1$ 时,术语 t 为类 c 的负依赖测度;当 $R_{t,c}=0$ 时,术语 t 和类 c 无依赖关系。

2.3　术语-术语关系映射

长期关系映射是深入学习过程,其目的是对术语列表中每个术语关系进行计算。在类关系映射中,术语与类间存在加权,但不知如何将这些术语与类别关联。因此,需进一步发现术语关系映射并计算这些条款间的加权关系。计算并创建 2 个术语间的直接关系,如图 4(b)所示。

为衡量术语-术语关系,首先在每个类中选择特定数量条款。在真实情况下,确定阈值 k 进行术语-类别计算:

步骤 1　对于 m 种类别,每个类别含有最高等级 k 的正项术语:

$$\begin{cases}Y=\{y_1,y_2,\cdots,y_m\}\\ v_i=\left\{\begin{matrix}(t_1,\chi^2_{t_1,c_i},R_{t_1,c_i}),(t_2,\chi^2_{t_2,c_i},R_{t_2,c_i}),\cdots,\\(t_k,\chi^2_{t_k,c_i},R_{t_k,c_i})\end{matrix}\right\}\\ R_{t_j,c_i}=O_{t_j,c_i}/E_{t_j,c_i}>1\end{cases} \tag{6}$$

其中,R_{t_j,c_i} 表示术语 $t_j(t_j\in T,T=\{t_1,t_2,\cdots,t_k\})$ 对于类别 $c_i(c_i\in C,C=\{c_1,c_2,\cdots,c_m\})$ 的依赖测度,为正值。

步骤 2　如果正依赖测度术语($R_{t,c}>1$)数量小于阈值 k,那么选择类别中所有的正值术语构建术语组。2 个术语 t_a 和 t_b 的共现频率可通过修改式(4)和式(5)进行计算:

$$\chi^2_{t_a,t_b}=\sum_{i\in\{t_a,-t_a\}}\sum_{j\in\{t_b,-t_b\}}\frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}} \tag{7}$$

$$R_{t_a,t_b}=\frac{O_{t_a,t_b}}{E_{t_a,t_b}} \tag{8}$$

其中,$\chi^2_{t_a,t_b}$ 表征数据集的文件中术语 t_a 和 t_b 间的依赖测度;R_{t_a,t_b} 用来区分正负测度;经过术语-术语关系映射和归一化操作后,可得到 2 个 $k\times k$ 的术语-术语依赖矩阵,包含 $n\chi^2_{t_a,t_b}$ 和 R_{t_a,t_b}。

上述独立表示的术语可转换成有向图 $G=(V,A)$,其中,V 为选定的术语集合;$V=\{t_1,t_2,\cdots,t_{k-1},t_k\}$;$A$ 为术语间的直接关系权重,形式为:

$$A=\{(t_i,t_j,n\chi^2_{t_i,t_j})\mid i\in\{1,2,\cdots,k\},j\in\{1,2,\cdots,k\}\} \tag{9}$$

根据 R_{t_i,t_j} 的取值,可判定 t_i,t_j 是否在有向图 G 中进行表示。当 $R_{t_i,t_j}>1$ 时,有向图 G 中术语 t_a 和 t_b 间存在可见连接。当 $R_{t_i,t_j}<1$ 时,则表示在有向图

G 中术语 t_a 和 t_b 间不存在直接连接。图 5 为 $k=4$ 以及所有 $R>1$ 情况下的有向连接图。

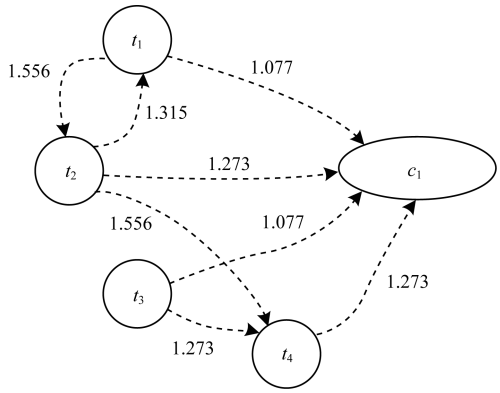


图 5 有向连接图

图中节点数值定义为：

$$v_{c_1}=\left\{\begin{matrix}(t_1,0.437,1.764),(t_2,0.346,1.579),\\(t_3,0.476,1.667),(t_4,0.418,1.483)\end{matrix}\right\} \tag{10}$$

3 模型创建及学习过程

3.1 概念聚类

概念聚类是将语义相似的术语分配到一个语义组的过程。在上述步骤中创建的本体图形作为概念聚类的基础输入。思路是将高权重关系的术语组合为一个子图，将低权重关系的术语语义剔除，构建新子图。簇是自动创建的，无需明确定义所需创建的集群数。首先，选取连接权重最大 e_x 的 2 个顶点 t_a 和 t_b 进行组合并作为初始聚类。然后，选择剩余最大权重边缘 e_y 的 2 个顶点 t_c 和 t_d，如果下一个选定顶点与现有集群任何顶点连接，则将顶点放入该集群；否则，构建新的聚类集群。

算法 1 概念聚类算法

```
For every DOG do
    选择最高权重边缘 e_x = {(t_a ,t_b ,R_{t_a,t_b})};
    创建首个包含 t_a 和 t_b 的概念聚类;
    For every e in V do
        选择最高权重边缘 e_x = {(t_c ,t_d ,R_{t_c,t_d})};
        If t_c ,t_d 在已有类别中 then
          将 t_c ,t_d 放在已有聚类中;
        Else
          创建新类别包含 t_c ,t_d ;
        Endif
    Endfor
Endfor
```

因此，概念聚类可创建二分类关系。概念层次第一层用于创建术语-类别关系映射，进一步聚集所有与类相关的术语，形成二层层次结构。每个集群创建一个与之相关的类作为父代，与之相关的所有术语作为子代，在 DOG 中形成三级分类层次概念。

3.2 领域本体图构建

在 KLSeeker 中，定义一组本体图概念模型，其中由一组术语及其相互关系构建概念。

定义 1（DOG） 在 KLSeeker 系统中，DOG 是与特定领域相关的本体图，可定义为：

$$OG_d=(T,F,H,R,C,A) \tag{11}$$

其中，d 为本体图关联域；T 为 OG_d 中术语 t_i 的集合；H 为 T 的分类关系集合；R 为 t_i 和 t_j 间的关系集合，$t_i,t_j\in T$；C 为聚类 $t_1,t_2,\cdots,t_n$ 集合，$t_1,t_2,\cdots,t_n\in T$；A 为 R 的关系特征。

在本体学习过程中，通过对中文语料库进行分类获得 DOG。手动过程包括：定义初始术语列表（可从已有的字典中获得），定义和映射类别字函数（也可已有的字典中获得）并标记概念群。自动过程包括：领域术语提取，术语关系提取以及概念聚类提取。

算法 2 DOG 生成算法

输入 T,F,H,R,C,A

输出 DOG 本体图概念模型

```
获取术语列表 T = {t_1 ,t_2 ,⋯ ,t_k };
For t_i ∈ T do
  在 DOG 中生成节点 n_i ;
  根据术语名称进行节点标记;
  根据术语类别测度 χ²_{t,d} 进行节点权重赋值;
Endfor
获得术语列表 T 依赖测度并令 θ 为 DOG 中最小测度值
For 每个术语-术语关系映射 do
  If t_a ,t_b 的测度值≥θ then
      在 DOG 中生成 t_a ,t_b 边缘 e_i ;
      设置 e_i 权重为 χ²_{t_a,t_b};
  Endif
Endfor
移除 DOG 中所有未连接权重;
  End
```

3.3 文档本体图生成

DocOG 是另一种类型的本体图，用来表示单一文本文件内容。传统信息检索系统通常用术语向量表示文档。在 KLSeeker 系统中，采取本体图模型来表示文本文档的内容。此外，该 DocOG 也可描述文件的更多信息，如特定领域的相关知识，可通过匹配文本文件，获得更多的 DOG 知识相关域。

文本文档匹配的目的是提取文档中域的更多知识。文本文档通常用一个术语列表表示（加权项向量）。如果它们有相同的术语交集，则可构建术语向量与 DOG 的映射，如图 6 所示。

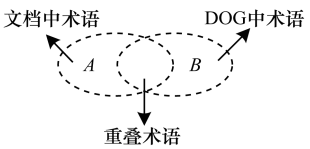


图 6 文档术语和 DOG 的重叠映射

此过程可将文档与特定域进行关联,可用来生成原文件的 DocOG,同时也可衡量文档与 DOG 的匹配相似性。

算法 3　DocOG 生成算法

输入　文档并生成 $DOG\text{-}OG_d$

输出　$DocOG(OG_{doc})$

获得文档内容术语列表 $T=\{t_1,t_2,\cdots,t_k\}$;

获得术语集 T_d 和关系集 R_d;

将文档转换为权重术语向量:

$V_{doc}=\{(t_1,w_{t_1}),(t_2,w_{t_2}),\cdots,(t_k,w_{t_k})\}$;

对于 DocOG(OG_{doc})中术语 T_{doc} 和关系集 R_{doc} 进行本体图匹配;

for $t_i \in V_{doc}$ do

　If t_i 存在于 T_d 中 then

　　For R_d 中的每个关系术语 t_j do

　　将 t_i,t_j 添加到术语集 T_{doc} 中;

　　将 t_i,t_j 关系添加到关系集 R_{doc} 中;

　　计算 t_i,t_j 权重 $w_{t_i,t_j}=w_i$,w_j;

　　Endfor

Endif

　Endfor

根据 T_{doc} 和 R_{doc} 生成本体图;

End

3.4　本体文本分类

本文通过测量 DocOG 和 DOG 间的相似性对本体图文本进行分类。当 DocOG 与多个 DOG 进行比较时,得分最高的 DOG 是文档所属领域。

算法 4　本体文本分类算法

输入　文档 DocOG 和领域知识 DOC

输出　基于相似等级的类别标签

1)对于每个领域产生相应的 DOG,如果领域中有 m 个类别,则创建 m 个 DOG;

2)对于每个文档 DOG 产生数量为 m 的 DocOG 关系权重;

3)基于所有关系权重计算每个 DocOG 的评价值;

4)选择最高评价值领域作为分类结果。

4　实验与结果分析

4.1　DOG 和 DocOG 实例计算比较

在本节中将给出一个实例来说明 DOG 和 DocOG 的匹配过程。利用表格形式表示域 d 中的 DOG,具体如表 1 所示。$DocOG_{d_1}$ 和 $DocOG_{d_2}$ 分别表示 2 个不同的文档 d_1 和 d_2,具体的关系信息如表 2、表 3 所示。

表 1　DOG_d 关系表

节点	A	B	C	D	E
A	0.10	0.01	0.02	0.03	0.04
B	0.01	0.10	0.02	0.03	0.04
C	0.01	0.02	0.10	0.03	0.04
D	0.01	0.02	0.03	0.10	0.04
E	0.01	0.02	0.03	0.04	0.10

表 2　DOG_{d_1} 关系表

节点	A	B	C	D	E
A	0.50	0.05	0.100	0.150	0.20
B	0.05	0.25	0.050	0.075	0.10
C	0.05	0.05	-	0.075	-
D	0.05	0.05	0.075	0.250	0.10
E	0.05	0.05	-	0.100	-

表 3　DOG_{d_2} 关系表

节点	A	B	C	D	E
A	-	-	-	0.225	0.10
B	-	-	-	0.225	0.20
C	-	-	-	0.225	0.10
D	0.075	0.15	0.225	0.750	0.30
E	0.025	0.05	0.075	0.300	0.25

首先,计算 DOG_{d_1} 和 DOG_{d_2} 关系表中所有数据总和与 DOG_d 表中所有数据总和的差作为每个 DocOG 的文档评价值:

$$score(DocOG_{d_1},DOG_d)=1.425 \tag{12}$$

$$score(DocOG_{d_2},DOG_d)=2.275 \tag{13}$$

然后,可获得 DocOG 和 DOG 间的相似性:

$$sim(DocOG_{d_1},DOG_d)=\frac{score(DocOG_{d_1},DOG_d)}{score(DOG_d)}=\frac{1.425}{5}=0.285 \tag{14}$$

$$sim(DocOG_{d_1},DOG_d)=\frac{score(DocOG_{d_1},DOG_d)}{score(DOG_d)}=\frac{2.275}{5}=0.455 \tag{15}$$

4.2　实验数据集

本文选取不同来源以及性质的测试集:中文网页测试集(PKU),搜狗分类测试集(Sogou)及复旦大学分类测试集(FD)。测试集信息如表 4 所示。

表 4　测试集信息

指标	PKU	Sogou	FD
文档数	15 511	17 910	19 636
类别数	12	9	20
长度/bit	6 ~ 307 152	118 ~ 194 195	78 ~ 67 218
长度均值/bit	4 813 ~ 67 956 324	2 816 ~ 18 563 254	1 007 ~ 54 652 589
文档数量	139 ~ 2 765	1 986	51 ~ 3 218
文档数量均值	1 287 ~ 865 423	0 ~ 1 986	983 ~ 1 156 284

在表 4 中,PKU 测试集的文档长度取值范围较大,Sogou 测试集最小,FD 测试集最大。FD 测试集的每个类别中文档数量变化范围最大,而其类别所含文档均值数量最少,Sogou 测试集类别所含文档数量相等,均为 1 986。

4.3 性能对比分析

本节实验分为两部分:1)在不同阈值选取情况下,比较数据集的分类精度;2)本文算法与其他算法的对比情况。算法评价指标[13]:

1)分类精度。通过计算正确检索文件占整个检索结果集的比例,测量检索模型精度。

$$precision = \frac{TP}{TP + FP} \tag{16}$$

2)召回率。通过计算正确检索文档所占应被检索文件的比例,获取召回率。

$$recall = \frac{TP}{TP + FN} \tag{17}$$

3)F 测量值。测量精度和召回率的调和平均数。

$$f\text{-}measure = \frac{2 \times precision \times recall}{precision + recall} \tag{18}$$

在上述定义中,TP 为被检索文件数;FP 为未被检索文件数;FN 为被检索为不相关的文件数。

对比算法选取文献[13-15]算法,其中,文献[13]算法通过对大型文本语料库进行非负张量分解,实现本体概念语义匹配;文献[14]算法基于阿拉伯语数据集进行文本分类技术设计。这 2 种算法都是针对非中文语言进行设计。文献[15]算法是基于表观语义和潜在性的中文文本分类方法,其是专门针对中文的文本分类算法。设置阈值 $\theta = 0$,实验对比结果如表 5 所示。

表 5 文本分类精度对比 %

测试集	算法	分类精度	召回率	F 测量值
PKU	文献[13]算法	86.3	75.6	80.6
	文献[14]算法	88.5	81.4	84.9
	文献[15]算法	93.4	91.5	92.4
	本文算法	96.7	95.4	96.0
Sogou	文献[13]算法	84.3	85.2	84.7
	文献[14]算法	89.7	83.4	86.4
	文献[15]算法	92.5	93.6	93.0
	本文算法	95.8	97.8	96.8
FD	文献[13]算法	89.6	82.3	85.8
	文献[14]算法	93.4	85.4	89.2
	文献[15]算法	95.9	90.5	93.1
	本文算法	98.1	92.3	95.1

由表 5 实验对比结果可知,与文献[13-15]算法相比,本文算法的中文文本分类精度、召回率以及 F 测量值指标均最优。由于文献[13-14]算法分别针对英语和阿拉伯语,因此其应用在中文文本分类中的效果并不理想。文献[15]算法是针对中文语料库进行的算法设计,其算法效果优于文献[13-14]算法,但差于本文算法。

4.4 阈值参数对文本分类精度的影响

阈值参数 θ 的作用是过滤依赖测度低于该值的边缘,因此,较高的阈值 θ 将减少本体图的边缘数量,从而进一步减少本体图生成规模。在实验中,令 $\theta \in [0,1]$,不同阈值参数 θ 对本文算法分类精度、召回率以及 F 测量值指标在 PKU,Sogou 以及 FD 测试集上的实验结果如图 7 所示。

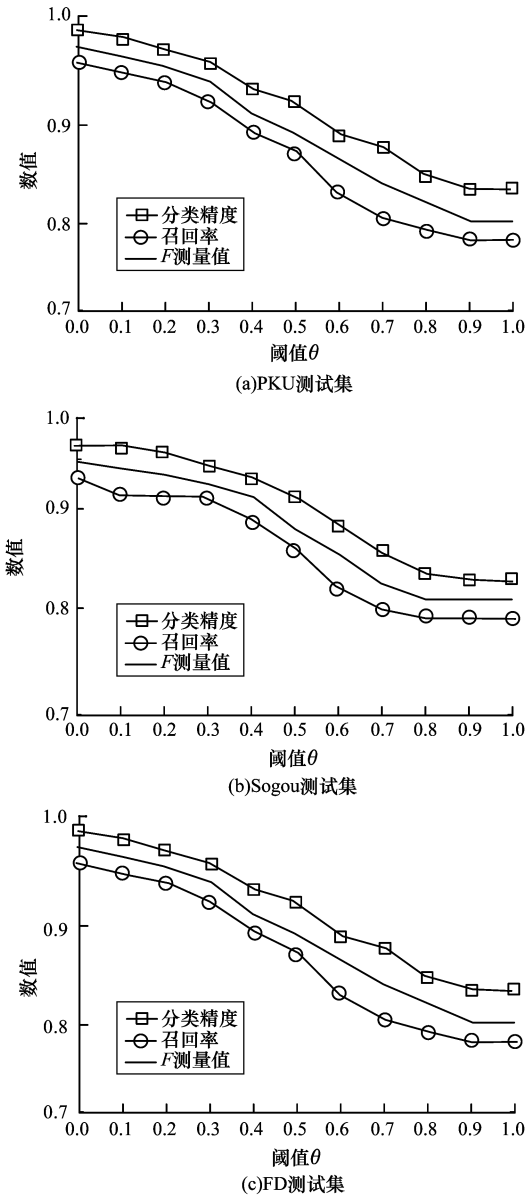


图 7 阈值参数对文本分类精度的影响

由实验数据可知,随着阈值参数 θ 取值的增大,本文算法分类精度、召回率以及 F 测量值指标均呈现下降态势,说明随着阈值参数的增大,算法本体图

结构规模降低,有助于算法计算效率的提升,但同时会降低算法的分类精度,因此,如何选取阈值参数 θ 需要根据实际进行选择。

5　结束语

本文基于半监督概念聚类的领域本体图对中文文本分类问题进行研究,解决英语词消歧方法无法应用于中文文本分类的问题,利用领域本体图结构模型、HowNet 字典及 KLSeeker 本体中文文本分类算法等实现中文文本的精确分类。但由于本体图是对应用的简化,概念间的关系类型不能通过半监督方式进行本体学习,此外,目前仍缺乏有效的本体论验证来衡量所产生 DOG 的完整性,因此下一步将对上述内容进行深入研究。

参考文献

[1] 杨月华,杜军平,平　源. 基于本体的智能信息检索系统[J]. 软件学报,2015,26(7):1675-1687.

[2] 石　敏,赵文栋,张　磊. 一种基于本体划分的语义 Web 服务发现算法[J]. 计算机工程,2014,40(2):175-179.

[3] Uchibayashi T, Apduhan B O, Shiratori N. A Domain Specific Sub-ontology Derivation End-user Tool for the Semantic Grid[J]. Telecommunication Systems, 2014, 55(1):125-135.

[4] Mallik A, Chaudhury S. Acquisition of Multimedia Ontology: An Application in Preservation of Cultural Heritage [J]. International Journal of Multimedia Information Retrieval,2012,1(4):249-262.

[5] Lehmann J, Chan M, Bundy A. A Higher Order Approach to Ontology Evolution in Physics[J]. Journal on Data Semantics,2013,2(4):163-187.

[6] Ali F, Kim E K, Kim Y G. Type-2 Fuzzy Ontology-based Opinion Mining and Information Extraction: A Proposal to Automate the Hotel Reservation System[J]. Applied Intelligence,2015,42(3):481-500.

[7] 崔　恺,凌兴宏,姚望舒. 一种基于马尔科夫网的本体匹配算法[J]. 计算机工程,2013,39(11):209-213.

[8] Sellami Z, Camps V, Aussenac-Gilles N. DYNAMO-MAS: A Multi-Agent System for Ontology Evolution from Text[J]. Journal on Data Semantics,2013,2(2):145-161.

[9] Garanina N O, Sidorova E A. Ontology Population as Algebraic Information System Processing Based on Multi-agent Natural Language Text Analysis Algorithms [J]. Programming and Computer Software, 2015, 41 (3): 140-148.

[10] Getman A P, Karasiuk V V. A Crowdsourcing Approach to Building a Legal Ontology from Text[J]. Artificial Intelligence and Law,2014,22(3):313-335.

[11] Javed M, Abgaz Y M, Pahl C. Ontology Change Management and Identification of Change Patterns[J]. Journal on Data Semantics,2013,2(2):119-143.

[12] Kumazawa T, Kozaki K, Matsui T. Initial Design Process of the Sustainability Science Ontology for Knowledge-sharing to Support Co-deliberation [J]. Sustainability Science,2014,9(2):173-192.

[13] Anisimov A V, Marchenko O O, Vozniuk T G. Determining Semantic Valences of Ontology Concepts by Means of Nonnegative Factorization of Tensors of Large Text Corpora [J]. Cybernetics and Systems Analysis,2014,50(3):327-337.

[14] Khorsheed M S, Al-Thubaity A O. Comparative Evaluation of Text Classification Techniques Using a Large Diverse Arabic Dataset[J]. Language Resources and Evaluation,2013,47(2):513-538.

[15] Chen Yewang, Wang Jiongliang, Cai Yiqiao. A Method for Chinese Text Classification Based on Apparent Semantics and Latent Aspects[J]. Journal of Ambient Intelligence and Humanized Computing, 2015, 6 (4): 473-480.

编辑　陆燕菲

(上接第 180 页)

[9] Sánchez J, Perronnin F, Mensink T, et al. Image Classification with the Fisher Vector: Theory and Practice[J]. International Journal of Computer Vision, 2013,105(3):222-245.

[10] Sydorov V, Sakurada M, Lampert C H. Deep Fisher Kernels——End to End Learning of the Fisher Kernel GMM Parameters [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2014: 1402-1409.

[11] Sánchez J, Redolfi J. Exponential Family Fisher Vector for Image Classification[J]. Pattern Recognition Letters, 2015,59(7):26-32.

[12] Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection [C]//Proceedings of IEEE Conference on Computer Vision & Pattern Recognition. Washington D. C., USA: IEEE Press, 2005: 886-893.

[13] Ma Ding, Pang Xin, Qu Mingzhi, et al. A Target Tracking Method Combining HOG and Shape Context[J]. Journal of Computational Information Systems, 2014,10(8):7053-7060.

[14] 王知音,禹　龙,田生伟,等. 基于栈式自编码的水体提取方法[J]. 计算机应用,2015,35(9):2706-2709.

[15] Bengio Y, Lamblin P, Popovici P, et al. Greedy Layer-wise Training of Deep Networks[J]. Advances in Neural Information Processing Systems,2007,19(1):153-160.

[16] 李海峰,李纯果. 深度学习结构和算法比较分析[J]. 河北大学学报(自然科学版),2012,32(5):538-544.

[17] Li Feifei, Fergus R, Perona P. Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories [J]. Computer Vision & Image Understanding,2007,106(1):59-70.

[18] Yann Lecun, Cortes C, Christopher J C. The MNIST Ddatabase of Handwritten Digits [EB/OL]. (2013-10-30). http://yann.lecun.com/exdb/mnist/.

编辑　索书志