基于百科词条的本体概念聚类方法研究

于 娟，曹 晓

(福州大学经济与管理学院 福州 350116)

【摘要】该文面向本体关系集合的自动构建，提出一种基于百科词条的本体概念聚类方法，用于发现领域概念之间的语义关系。在给定领域本体概念集合的条件下，该方法首先获取相关的百科词条并建立每一概念的向量模型，然后根据距离判别法进行概念聚类，得到概念间的相近关系。采用该方法对3个领域中的领域概念集合进行聚类，实验结果表明，该文方法比传统聚类算法有更好的聚类结果，有助于概念间关系的自动获取和领域本体自动构建。

关 键 词 概念聚类；距离判别法；百科词条；本体概念；本体学习

中图分类号 TP181 **文献标志码** A doi:10.3969/j.issn.1001-0548.2017.03.026

Ontology Concepts Clustering Based on Encyclopedia Entries

YU Juan and CAO Xiao

(School of Economics and Management, Fuzhou University Fuzhou 350116)

Abstract To build the set of ontology relations automatically, this paper presents a preliminary study on a concept clustering method based on encyclopedia entries of obtaining semantic relations among concepts. Given a set of domain ontology concepts, this method clusters concepts in 3 steps: 1) obtaining encyclopedia entries; 2) modeling each of the concepts into a vector; 3) clustering concepts using the distance discrimination method. Clustering experiments on 3 sets of domain ontology concepts demonstrate that the proposed method shows better results compared with classical clustering methods and has good potentials for identifying related concepts automatically in the ontology building tasks.

Key words concept clustering; distance discrimination method; encyclopedia entries; ontology concept; ontology learning

作为语义Web和知识管理系统的关键基础，本体描述某个领域甚至更广范围内的概念以及概念之间的关系，使得这些概念和关系在共享的范围内具有共同认可的、明确的、唯一的定义，以供人之间以及机器之间进行交流[1]。目前，本体在语义检索、知识管理和人工智能等相关领域得到了广泛的理论和应用研究。

本体学习是采用机器学习方法(半)自动构建本体的过程。根据学习的本体对象不同，本体学习主要包括概念学习、关系学习和公理学习。其中，关系学习试图采用计算机(半)自动地快速地发现概念间关系。在这个信息迅速增长的时代，新概念层出不穷，概念间关系发生着变化，因此关系学习是当前本体研究的重点和热点之一。

本文研究一种基于百科词条的本体概念聚类方法，用于支持自动发现概念间的关系。该方法首先依据百科词条建立概念的向量模型，然后根据距离判别法进行概念聚类，进而获取概念间的相关关系。

1 研究现状

聚类是将对象的集合分成相似的对象类的过程[2]。概念聚类是将概念集合分成相似的几类的过程。由于同一类的术语会在相同的上下文出现，所以可以通过聚类算法将上下文相似的术语进行聚类，进而发现概念间关系。概念聚类算法主要分为：划分聚类、层次聚类、形式概念分析和基于图结构的聚类。

1) 划分聚类。首先，构建对象的k个划分，然后采用迭代重定位技术，尝试通过对象在组间移动来改进划分。典型的划分方法有k均值(k-means)和k

收稿日期：2015－10－19；修回日期：2016－05－31
基金项目：国家自然科学基金(71201032)；福建省社会科学规划项目(FJ2016C044)
作者简介：于娟(1981－)，女，博士，副教授，主要从事领域本体、信息管理方面的研究.

中心点，一般都是对*k*-means算法的优化。文献[3]在计算聚类中心时，先删掉向量与平均向量相差超过10%的术语，再重新计算每个类的平均向量优化聚类中心。文献[4]依据类中元素分布计算类中聚集程度最大的*p*个概念，将距离这*p*个概念的平均向量最近的概念作为类的新中心，优化聚类中心。

2) 层次聚类。对给定概念集合进行层次的分解，构造一棵聚类树。层次聚类算法分为凝聚层次聚类和分裂层次聚类。文献[5]通过内部和外部凝聚层次聚类进行概念聚类。文献[6]对层次聚类算法适应性进行改造，通过计算层次的耦合−内聚比，计算类数目的分布密度。文献[7]计算最小增加值或最大减少值作为概念层次聚类的合并策略。

3) 形式概念分析。使用二元关系来表示领域中的形式背景，从形式背景中抽取概念层，即概念格，通过概念格结构将对象分层。文献[8]构造模糊概念格，对对象进行模糊概念聚类。文献[9]采用模糊*k*-means聚类算法约简概念格。

4) 基于图结构的聚类。文献[10]提出遍历树的蚂蚁聚类算法对术语进行聚类，用标准化的谷歌距离和Wikipedia测量术语间距离和相似度。文献[11]依据词性树路径长度、术语词汇相同词、连续词、开始结束词和术语概念层次树路径长度计算概念相似度，采用SOM自组织神经网络进行概念聚类。

前述研究中，聚类算法大多需要根据实验或者经验来设定阈值，而阈值对聚类结果的影响很大。阈值设置过大，可能丢失有趣的关联；阈值设置过小，可能产生大量的弱相关的交叉支持模式关联。并且，阈值的设定存在不确定因素，设置适当的阈值较为困难。另一方面，根据语境构建概念模型以计算概念间相似度时，少有研究将整个语料作为共现窗口。因此，本文以百科词条为语料，研究了一种无监督的概念聚类方法。

2　基于百科词条的概念聚类

本文提出一种基于百科词条的本体概念聚类方法。该方法的输入是：领域本体的概念集合、概念的百科词条；输出是：候选本体关系集合。方法流程图如图1所示。

对图1中各个模块的说明：

1) 领域概念集合即领域本体的概念集合，是领域专有概念的全集，是构建领域本体的基础。领域概念集合可以由领域专家人工给出，也可借助一些机器学习方法采用人机结合的方式获得，例如，文献[12]提出或改进的一系列领域概念学习方法。

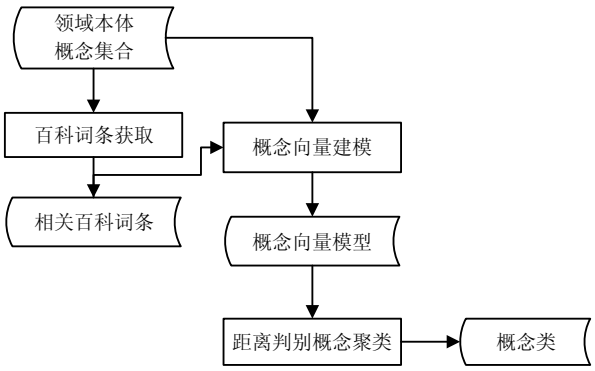

图1　基于百科词条的概念聚类方法流程图

2) 百科词条获取模块。对概念集合中的每一个概念查找其百科词条并进行预处理，获取其中的文本信息。

3) 相关百科词条是领域概念的百科词条，可以是百度百科、维基百科等词条。

4) 概念向量建模模块。对百科词条进行分词与信息熵过滤，统计共现词语及其词频建立领域概念的量化模型。

5) 概念向量模型表示概念的向量模型，其中每一个分量是一个共现词的词频。

6) 距离判别概念聚类模块。采用马氏距离计算概念间距离，重心距离计算概念到类中心的距离，经过多次迭代直至聚类结果不再改变。

7) 概念类是概念聚类所得的聚类结果。

2.1　概念向量建模

概念向量建模的输入是：领域本体的概念集合、概念的百科词条；输出是：概念向量模型。该模块对于每一个领域概念，通过对概念百科词条的预处理，计算信息熵过滤词语，所得概念向量模型中的每个分量为过滤后词语的词频。

该模块首先将概念的百科词条进行语料预处理；然后计算每个概念的共现词语 w_k 和词频 $f_k(1\leqslant k\leqslant n_i)$。然后使用左右信息熵对共现词语进行过滤，并选择词频最高的前*n*个词语。左、右信息熵公式如下所示：

$$\mathrm{LE}(w)=-\sum_{l\in L}P(lw\,|\,w)\log_2 P(lw\,|\,w)$$

$$\mathrm{RE}(w)=-\sum_{r\in R}P(wr\,|\,w)\log_2 P(wr\,|\,w)$$

式中，*l*是词语*w*的左邻接字；*r*是*w*的右邻接字；$P(lw\,|\,w)$ 表示在出现词语*w*的情况下，*w*的左邻接字是*l*的条件概率；$P(wr\,|\,w)$ 表示在出现词语*w*的情况下，*w*的右邻接字是*r*的条件概率；$\mathrm{LE}(w)$ 为词语*w*

的左信息熵； RE(w) 为词语w的右信息熵；左、右信息熵越大，则w越独立。综合考虑左右信息熵，得到如下信息熵公式：

$$\text{Entropy}(w)=\text{LE}(w)\times\text{RE}(w)$$

对百科词条语料中不独立的词语进行过滤，最后每个概念的概念向量模型表示为：

$$((w_1,f_1),(w_2,f_2),\cdots,(w_n,f_n))^{\text{T}}$$

该算法建立的概念向量模型的时间复杂度约为$O(n\times(m+k))$，其中，n为领域概念的数量，m为词语数量，k为过滤后的词语数量。

2.2 距离判别概念聚类

距离判别概念聚类的输入是：2.1节所得到的概念向量模型；输出是：概念类，即存在关系的概念的集合。本模块采用马氏距离计算概念间的距离，采用重心距离计算概念到类中心的距离。为叙述方便，先定义距离判别法所用到的数学符号和公式。

设有k个类(cluster)$c_1,c_2,\cdots,c_k$，它们的均值分别为$\mu^{(1)},\mu^{(2)},\cdots,\mu^{(k)}$，方差阵分别为$\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2,\cdots,\boldsymbol{\Sigma}_k$($\boldsymbol{\Sigma}_a>0$，$a=1,2,\cdots,k$)，$\boldsymbol{x}_{(1)}^{(a)},\boldsymbol{x}_{(2)}^{(a)},\cdots,\boldsymbol{x}_{(n_a)}^{(a)}$是来自类$c_a$的容量为$n_a$的概念向量($a=1,2,\cdots,k$)。类$c_a$的均值$\mu^{(a)}$和方差阵$\boldsymbol{\Sigma}_a$($a=1,2,\cdots,k$)估计公式如下：

$$\hat{\mu}^{(a)}=\bar{\boldsymbol{x}}^{(a)},\quad \hat{\boldsymbol{\Sigma}}_a=\frac{1}{n_a-1}\boldsymbol{L}_a$$

式中，$\bar{\boldsymbol{x}}^{(a)}$和$\boldsymbol{L}_a$分别是类$c_a$($a=1,2,\cdots,k$)的概念向量均值和离差阵：

$$\bar{\boldsymbol{x}}^{(a)}=\frac{1}{n_a}\sum_{i=1}^{n_a}\boldsymbol{x}_{(i)}^{(a)},\quad \boldsymbol{L}_a=\sum_{i=1}^{n_a}(\boldsymbol{x}_{(i)}^{(a)}-\bar{\boldsymbol{x}}^{(a)})(\boldsymbol{x}_{(i)}^{(a)}-\bar{\boldsymbol{x}}^{(a)})^{\text{T}}$$

为了描述概念间的相似程度，采用马氏距离计算概念之间的距离：

$$D(\boldsymbol{x},\boldsymbol{y})=\sqrt{(\boldsymbol{x}-\boldsymbol{y})^{\text{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{y})}$$

式中，$\boldsymbol{x}$和$\boldsymbol{y}$表示概念向量；$\boldsymbol{\Sigma}^{-1}$是类方差矩阵的逆矩阵。

采用重心距离计算概念与类中心的距离，计算公式如下：

$$D(\boldsymbol{y},C)=D(\boldsymbol{y},\mu)$$

式中，μ是类C的均值；$\boldsymbol{y}$是任意一个概念向量。

距离判别概念聚类方法步骤如下：

1) 人为地将概念分为k个类$c_1,c_2,\cdots,c_k$，计算每个类c_a的均值$\mu^{(a)}$和方差阵$\boldsymbol{\Sigma}_a$($a=1,2,\cdots,k$)；

2) 采用马氏距离计算概念之间的距离，采用重心距离计算概念与类中心的距离；

3) 将距离类c_a近的概念归为该类，经过多次迭代直至判别结果不再发生变化，得到最终的聚类结果。

该算法形成概念聚类的时间复杂度约为$O(kmtn^2)$，其中，n为领域概念的数量，m为词语数量，k为类的数量，t为迭代次数。

3 实验分析

为了验证本文提出的基于百科词条的本体概念聚类方法的性能，采用该方法及多篇文献所使用的k-means聚类算法分别对3个领域本体的概念集合进行聚类。实验领域包括：电子商务领域、知识管理领域和管理信息系统领域。实验输入数据为由全国科学技术名词审定委员会[13]审定公布的领域概念集合，实验输出的概念聚类结果与领域专家手工划分的结果进行比较。

按照文献[14]和[15]设定本文的实验性能指标，定义某一个类c_i的类匹配度为：

$$m(c_i)=\max\left(\frac{|c_i\cap c_j'|}{|c_i\cup c_j'|}\right)$$

式中，$m(c_i)$表示c_i和c_j'之间的一致程度，$0\leqslant m(c_i)\leqslant 1$；$c_i$是由领域专家人工确立的领域概念主题划分中的一个主题，$i=1,2,\cdots,n$，n为主题个数；c_j'为由聚类方法自动得到的一个类，$j=1,2,\cdots,n$；$|c_i\cap c_j'|$为c_i和c_j'中相同概念的个数。

定义聚类结果C'的匹配度为：

$$m(C,C')=\sum_{i=1}^{n}m(c_i)P(c_i)$$

式中，C和C'分别表示领域专家人工确立的领域概念集合的主题划分和聚类方法得到的结果；$m(C,C')$反映了两种结果C和C'的一致程度，$0\leqslant m(C,C')\leqslant 1$，1代表完全一致，0代表完全不一致；$n$为领域专家人工确立的领域概念主题划分的主题个数；$P(c_i)$是$c_i$发生的概率。

图2～图4展示了本文方法及k-means对3个领域概念集合的聚类结果与领域专家人工划分结果的类匹配度。

表1采用4项经典的聚类性能指标对本文方法及k-means的聚类结果进行比较，并统计了3个领域的本文方法及k-means的聚类结果与领域专家人工确立的领域概念集合的主题划分的匹配度。

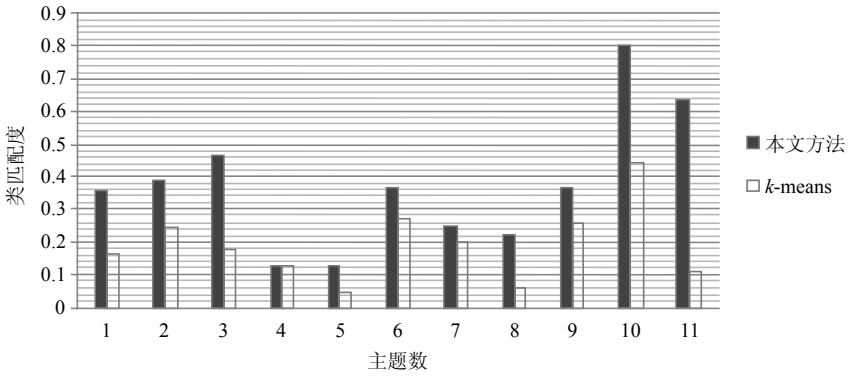

图2　电子商务领域概念聚类情况

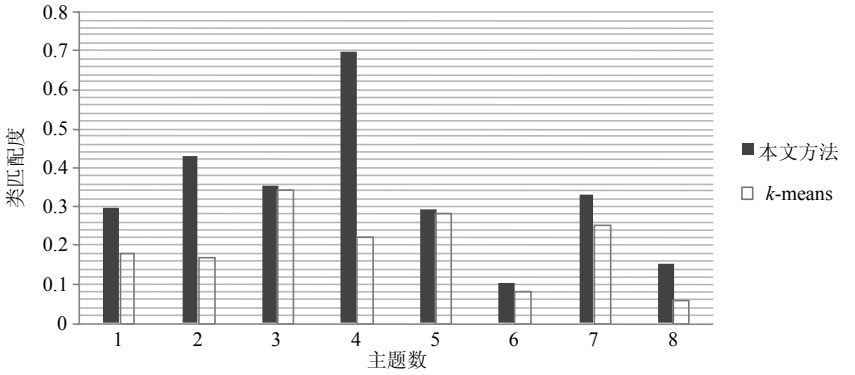

图3　知识管理领域概念聚类情况

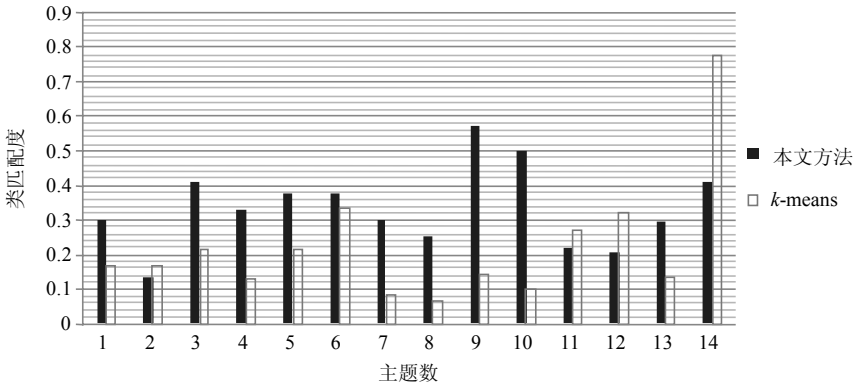

图4　管理信息系统领域概念聚类情况

表1　3个领域聚类结果比较

评价指标	电子商务领域		知识管理领域		管理信息系统领域	
	本文方法	*k*-means	本文方法	*k*-means	本文方法	*k*-means
候选关系数	888	2 179	621	745	1 616	3 324
准确数	258	357	195	152	350	539
关系数	698	698	580	580	1 149	1 149
Precision	0.291	0.164	0.314	0.204	0.217	0.162
Recall	0.370	0.511	0.336	0.262	0.305	0.469
F-Score	0.326	0.248	0.325	0.230	0.254	0.241
RI	0.815	0.626	0.773	0.714	0.827	0.716
匹配度	0.373	0.191	0.333	0.198	0.334	0.223

表1中设每个类发生的概率相同，“本文方法”一列的数据为使用本文提出的概念聚类方法所得到的本体关系，“*k*-means”一列的数据为使用*k*-means聚类算法得到的本体关系。在不影响比较结果的情况下，将表1中的数据四舍五入。对评价指标的说明如下：

1) 候选关系数表示自动学习所得到的候选本体关系的数目。

2) 准确数表示候选本体关系中被领域专家人工确立的属于同一主题的数目。

3) 关系数表示领域专家人工确立的本体关系集合中所有本体关系的数目。

4) Precision=准确数/候选关系数。

5) Recall=准确数/关系数。

6) F-Score是Precision和Recall的调和平均数。F-Score计算公式：

$$\text{F-Score}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision+Recall}}$$

7) RI (rand index)是随机一致性指标，用来度量聚类结果与领域专家人工确立的概念所属主题分布的相似度。RI计算公式：

$$\text{RI}=\frac{\text{TP+TN}}{\text{TP+FP+TN+FN}}$$

式中，TP表示同一类的概念被分到同一个类，即准确数；TN表示不同类的概念被分到不同类；FP表示不同类的概念被分到同一个类；FN表示同一类的概念被分到不同类。

从表1的比较结果可以看出，对3个领域本体的概念集合进行聚类时，本文方法比*k*-means总体聚类匹配度高而且准确率高，说明了本文方法的有效性。

4 结 束 语

本文提出了一种基于百科词条的本体概念聚类方法，用于支持本体关系的自动获取。在给定领域概念集合的情况下，该方法首先获取概念的百科词条并从中获取文本信息，然后进行分词和信息熵过滤，增加建立的概念向量模型，最后采用距离判别法进行概念聚类。该方法不必确定阈值，使聚类算法更加自动化。实验结果表明该概念聚类方法有较好的聚类结果。

今后的研究方向将是，改进概念向量建模过程中的词语选取方法以及向量建模算法，在保证准确率的基础上提高召回率。

参 考 文 献

[1] GRUBER T R. A translation approach to portable ontology specifications[J]. Knowledge Acquisition, 1993, 5(2): 199-220.

[2] HAN Jia-wei, KAMBER M, PEI Jian. Data mining: Concepts and techniques[M]. 3rd ed. Beijing: China Machine Press, 2012: 443-444.

[3] 徐德智, JUNAID. Cluster-Merge本体构造算法[J]. 计算技术与自动化, 2010, 59(3): 49-52.
XU De-zhi, JUNAID. An ontology learning based on documents clustering[J]. Computing Technology and Automation, 2010, 59(3): 49-52.

[4] 胡云飞. 本体学习中关系获取的研究[D]. 西安: 西安建筑科技大学, 2012.
HU Yun-fei. Research on relations acquisition of ontology learning[D]. Xi'an: Xi'an University of Architecture and Technology, 2012.

[5] LEUNG K W T, LEE D L. Deriving concept-based user profiles from search engine logs[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(7): 969-982.

[6] 何琳, 侯汉清. 基于统计自然语言处理技术的领域本体半自动构建研究[J]. 情报学报, 2009, 28(2): 201-207.
HE Lin, HOU Han-qing. Research on semi-automatic construction of domain ontology based on statistical NLP technique[J]. Journal of the China Society for Scientific and Technical Information, 2009, 28(2): 201-207.

[7] CHEN Shi-xi, WANG Hai-xun, ZHOU Shui-geng. Concept clustering of evolving data[C]//IEEE 25th International Conference on Data Engineering. Shanghai, China: IEEE Computer Society, 2009: 1327-1330.

[8] THO Q T, HUI S C, FONG A C M, et al. Automatic fuzzy ontology generation for semantic web[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(6): 842-856.

[9] KUMAR C A, SRINIVAS S. Concept lattice reduction using fuzzy *k*-means clustering[J]. Expert Systems with Applications, 2010, 37(3): 2696-2704.

[10] WONG W, LIU W, BENNAMOUN M. Tree-traversing ant algorithm for term clustering based on featureless similarities[J]. Data Mining and Knowledge Discovery, 2007, 15(3): 349-381.

[11] LEE C S, KAO Y F, KUO Y H, et al. Automated ontology construction for unstructured text documents[J]. Data & Knowledge Engineering, 2007, 60(3): 547-566.

[12] 于娟. 基于文本的领域本体学习方法及其应用研究[D]. 大连: 大连理工大学, 2010.
YU Juan. Learning domain ontologies from Chinese text corpora[D]. Dalian: Dalian University of Technology, 2010.

[13] 全国科学技术名词审定委员会. 全国科学技术名词审定委员会简介[EB/OL]. [2016-12-24]. http://www.cnctst.cn/.
China National Committee for Terms in Sciences and Technologies. An introduction of China national committee for terms in sciences and technologies [EB/OL]. [2016-12-24]. http://www.cnctst.cn/.

[14] 刘金岭. 基于《现代汉语语义分类词典》的文本聚类方法[J]. 情报杂志, 2010, 29(11): 170-173.
LIU Jin-ling. Text clustering method based on thesaurus of modern Chinese[J]. Journal of Intelligence, 2010, 29(11): 170-173.

[15] 张明卫, 刘莹, 张斌, 等. 一种基于概念的数据聚类模型[J]. 软件学报, 2009, 20(9): 2387-2396.
ZHANG Ming-wei, LIU Ying, ZHANG Bin, et al. Concept-based data clustering model[J]. Journal of Software, 2009, 20(9): 2387-2396.

编 辑 蒋 晓