领域本体构建中关系辅助判断技术研究*

张晓莹，张桂平，王裴岩
沈阳航空航天大学 知识工程研究中心，辽宁 沈阳 110136
E-mail: zxy050401109@yahoo.com.cn

摘 要：领域本体辅助构建方法逐渐成为领域本体构建研究的热点，其中如何辅助用户判断概念间关系是领域本体构建的重点。针对用户在无领域背景知识支撑时无法准确判断概念间关系的问题，本文考虑文本中概念间距离对该文本描述概念间关系的影响，采用改进的BM25相似度计算方法为用户提供参考文本，并提出基于概念最短距离的分类样本提取方法，然后使用KNN分类算法实现增量式的关系推荐。实验结果表明，该方法可以有效的辅助用户判断概念间关系。

关键词：领域本体；辅助构建；信息检索；KNN分类

Research on Assistant Relation Judgement Technology in Domain Ontology Construction

Zhang Xiaoying, Zhang Guiping, Wang Peiyan
Knowledge Engineering Research Center, Shenyang Aerospace University, Shenyang, 110136
E-mail: zxy050401109@yahoo.com.cn

Abstract: Assistant approach becomes a research focus in the construction of domain ontology. Among these research, how to assist user in judging the relation between concepts is an important topic. According to the problem that users are unable to judge the relation between concepts accurately without the support of domain context, this paper considered the distance between concepts which influences the text's description of relation between concepts, and applied improved BM25 similarity computation approach to provide reference text to user. This paper also proposed a classification sample extraction approach based on minimum distance between concepts, and adopted KNN classification algorithm to incrementally recommend relation. Experiment results show that the new method can effectively assist user in judging the relation between concepts.

Keywords: domain ontology; assistant construction; information retrieval; KNN classification

1 引言

在人工智能界，最早给出本体定义的是 Neches[1]等人，他们将本体定义为“给出构成相关领域词汇的基本术语和关系”。领域本体是表达领域概念及概念间关系的知识集合，在知识共享、信息检索等方面有着重要的应用价值。目前领域本体的构建方法主要分为手工构建、全自动构建和半自动构建，其中半自动构建也称辅助构建。由于手工构建本体会耗费大量人力物力，而全自动的本体构建方式又将产生大量噪音数据、抽取的概念关系松散使得可信度无法得到保证[2]。因此，目前领域本体构建的研究集中于如何实现机器与领域专家的相互辅助，即辅助构建。

领域本体构建包括概念获取、关系判断以及领域本体形式化表示。通过对领域本体辅助构建方法的研究发现，大多数的领域本体辅助构建工具都是通过对数据源预处理后使用算法库中的本体学习算法获取领域本体，并将结果作为候选领域本体呈现给用户，用户评价和判断该结果，并将最终结果添加到领域本体库中[3][4][5]。这种先批量构建领域本体，后人工检查的方法使得机器操作与人工判断相互脱节。领域本体涉及概念广泛，结构关系复杂，若没有深厚的领域背景知识支撑，用户很难把握概念的具体含义及其与其他概念的关系，这便加大了本体构建的难度，因此，能否给用户提供领域背景知识是本体辅助构建过程中亟待解决的问题。再者，从目前领域本体辅

* 基金项目：国家自然科学基金资助项目（61073123）；辽宁省教育厅创新团队资助项目（LT2010084）。

助构建的研究方法来看，基于启发式规则的方法需要预先定义关系类别，对于新增类别没有很好的适应能力；而基于聚类的方法，只能获得相似概念间关系集合，不能给出确切的概念关系定义。

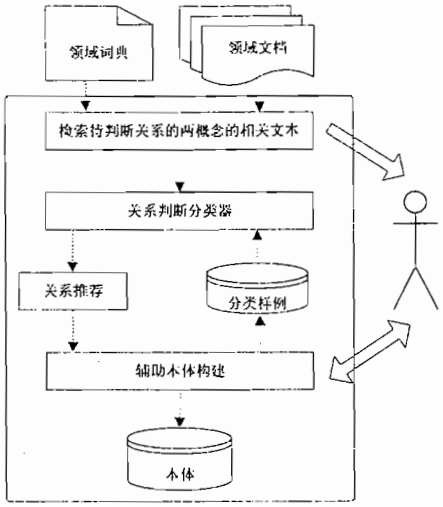

图1　本体构建中关系辅助判断基本框架

本文针对上述问题，首先采用基于改进的 BM25 算法的信息检索技术，从大量的领域文本库中获得包含两个待判断关系的概念的文本，作为用户理解领域概念和判断概念间关系的参考知识。其次，采用 KNN 分类算法为用户提供概念关系的推荐，利用 KNN 惰性学习的性质，使其对新增关系类别具有较好的自适应能力，同时，拥有随着关系推荐次数增多推荐效果越好的增量式学习能力。本文的方法框架如图 1 所示。

2　相关研究

构建领域本体的数据源一般分为三种：结构化数据、半结构化数据和非结构化数据，目前大多数的本体构建技术方面的研究主要集中在从非结构化数据中获取本体[6]。文献[7]利用领域词典构建航空领域本体；文献[8]抓取与领域相关的 Web 网页并从中提取纯文本作为本体进化的语料库源；阿姆斯特丹大学的 Wiefinga 等运用艺术和建筑叙词表（AAT）的受控词汇表描述古代家具本体等[9]。通过以上可以看出研究者们仅将词典、Web 网页或叙词表等作为获取本体的数据源，但却没有考虑从中检索有代表性的文本，提供给用户作为本体辅助构建的参考知识。

针对不同类型的数据源需要采用不同的本体构建方法，但通过对各种本体构建方法本质的研究，我们可将领域本体中概念间关系的判断方法分为：基于模板匹配的方法、基于关联规则的方法和基于统计学习的方法等。Hearst [10]等人在 1992 年提出了基于模板匹配的方法，该方法总结频繁出现的语言模式作为规则，若文本中词的序列匹配某个模式则可以识别出相应的关系。其中模板规则的获取需要领域专家参与，并且不能保证得到最好的模板规则。2000 年，Maedche [11]等人提出使用关联规则获取本体中的概念间关系，该方法能够较为准确的判断两个概念间是否存在关系，但属于哪种关系并不能明确的给出。2005 年，Kavalec[12]等人提出扩展关联规则的方法为概念间的关系赋予语义标签，其基本思想是使用经常出现在两个词附近的某个动词来表示这两个词的关系。该方法仅考虑了词频，没有考虑句子结构等其他因素，所以结果并不十分理想。D. Fature[13]采用基于分层的概念聚类方法，它的基本聚类器包含了一些词语固定搭配，这些搭配都由动词加介词的形式构成。L.Khan[14]等人使用聚类技术和 WordNet 从文本文档创建领域本体，创建过程自底向上，这种基于概念聚类的方法只能确定概念间的分类关系，而且关系的语义标签也难于确定。

本文使用领域词典与领域文档作为数据源，从中为用户检索判断概念间关系的参考文本，并使用 KNN 分类算法对概念间关系做出推荐。

3 概念关系辅助判断方法

本文使用两类数据源辅助构建领域本体：领域词典、领域文档。使用信息检索技术为用户提供概念间关系的文本参考，利用 KNN 算法推荐概念间的关系类别，用户根据系统提供的文本参考及关系推荐做出关系判断，并将此文本及其关系类别返回给系统，使得系统具有增量式的学习能力。本文考虑概念间距离对文本提供概念间关系能力的影响改进 BM25 算法，使用该算法计算得到的相似度值对检索结果排序，并根据概念间的距离提取 KNN 算法的分类样本。

3.1 基于信息检索技术的关系识别参考知识的获取

本文使用信息检索得到的文本作为判断概念关系的参考知识及 KNN 分类的样本，若有效的对检索结果排序，可以方便用户找到最相关的信息，并为 KNN 分类提供具有区分性的样本。本文针对数据源的特点制定排序准则并改进 BM25 算法对检索结果排序。

3.1.1 检索及其结果排序准则

领域词典与领域文档在内容及组织形式上均存在一定差异：领域词典是对领域概念的详细定义；领域文档描述的是领域内部的研究现状、新技术等。领域词典里某概念的释义项中若出现另一概念，则该释义项能够有力的提供这两个概念间关系的参考，而其他仅仅包含两概念的相关文本提供关系参考的能力则较弱。另外，两概念在文本中的距离也反映了该文本描述概念关系的能力，即两概念距离越近该文本描述概念关系的能力越强，根据这一分析制定如下排序准则。

假设待判断关系的两概念为 w_i、w_j，则在上述数据源中检索出的相关文本按以下准则排序：

1. 对于领域词典，若 w_i 或 w_j 出现在 w_j 或 w_i 的释义项中，则该释义项排序优先级最高。
2. 其余 w_i 与 w_j 共现的文档使用根据相对位置信息得到的相似度值由大到小排序。

根据概念相对位置信息改进的 BM25 相似度计算方法将在下一小节具体阐述。

3.1.2 改进的 BM25 相似度计算方法

BM25[15]是 1970 年由 S.E.Robertson 等在基于概率检索的框架上提出的，经典的 BM25 综合了特征在文本中的词频、平衡了文档长度等特征，其计算公式如下：

$$score(D,Q)=\sum_{i=1}^{n} IDF(q_i)\cdot\frac{f(q_i,D)\cdot(k+1)}{f(q_i,D)+k\cdot(1-b+b\cdot\frac{|D|}{avgdl})}$$

其中 Q 是待检索的向量$\{q_1,\ldots,q_n\}$，n 代表向量 Q 的检索词个数；D 是语料中的一个样本向量$\{w_1,\cdots,w_m\}$，m 是 D 的特征个数；$f(q_i,D)$是检索词 q_i 在 D 中出现的次数；$|D|$是文档 D 的长度；avgdl 是检索到的全部样本的平均长度；$IDF(q_i)$是 q_i 的反文档频度；k 和 b 是自由参数。

在数据源中检索出现两概念的相关文本，文本中两概念的相对位置是该文本描述两概念关系的一个重要因素，即两概念在文本中出现的位置越接近，则该文本描述两概念关系的能力越强。考虑图 2 两个均包含“飞机”和“机翼”的文本，可以较容易的看到文本 A 描述“飞机”与“机翼”间部件关系的能力比文本 B 更强。

A：“利用 飞机 的 机翼 产生 升力 的 原理 制造 一种 高性能 滑翔伞。”

B：“飞机 的 飞行 马赫数 指 飞机 的 空速 和 包围 飞机 的 大气 中 的 声速 之 比 。马赫数 的 概念 对 高亚 声速 机翼 的 后掠翼设计 起着 重要 的 作用 。”

图2 包含“飞机”与“机翼”的两个文本

由于本文的检索向量包含的检索词q_i在所有的样本向量中均出现，因此所有检索词q_i的IDF(q_i)的值均相同，故IDF(q_i)在改进的公式中不参与计算。本文在经典的BM25计算方法中添加了概念间的相对位置信息，改进后的公式如下：

$$score(D,Q)=\log(\frac{1}{\min_{Dis}(Q,D)+1})\cdot\sum_{i=1}^{n}\frac{f(q_i,D)\cdot(k+1)}{f(q_i,D)+k\cdot(1-b+b\cdot\frac{|D|}{avgdl})}$$

其中$\min_{Dis}$(Q,D)是Q中两检索词在向量D中的最短距离，即两词间的最少词语个数。

3.2 基于KNN算法的关系推荐

文献[16]提出了KNN分类算法，它是基于实例学习的基本算法之一，是一种非参数的分类技术，其样本集合可以调整，这些特性使其适用于本文的关系类别推荐。

3.2.1 基于概念最短距离的分类样本提取方法

从检索到的文档中提取具有区分性的文本作为KNN算法的分类样本，有利于提高分类效果与效率，进而取得较好的类别推荐性能。一般地，检索结果中位置靠前的文档包含概念关系的参考知识较多。另外，文档中包含两概念并使两概念距离最短的句子比整篇文档更有分类意义。因此，本文将排序位置因素与概念距离因素加入到分类样本提取的过程中，具体方法如下：

两概念<w_i,w_j>检索得到的前n篇文档集合为{D_1,… ,D_n}，每篇文档是句子的序列[S_1,… ,S_m]，其中m为句子个数。若w_i出现在S_i中，w_j出现在S_j中，并且使得w_i与w_j的距离最短（“距离”是指两概念间的词语数目），那么抽取文档D_k中的S_i和S_j（若S_i与S_j相同，则取一次）作为分类样本，即两概念<w_i,w_j>获得n篇分类样本。样本testDoc可形式化定义如下：

$$testDoc=\{<S_i,S_j>|\min_{Dis}(w_i,w_j),w_i\in S_i,w_j\in S_j,S_i\in D_k,S_j\in D_k\}$$

3.2.2 基于KNN算法的关系推荐方法

两概念< w_i,w_j >获得n个分类样本，KNN算法对每个样本均标识出关系类别，那么这两个概念的关系类别取决于哪个分类样本很难确定，本文采用下面的方法获得两概念的推荐关系。

KNN算法对两概念<w_i,w_j>获得的n个样本分类，每个样本均获得类别以及分值，将属于同一类别的分值相加作为该关系类别的可信度，按可信度值的大小给出< w_i,w_j>的关系类别推荐。

3.2.3 KNN算法的新增类别适应性与增量式学习

新增类别适应性是指学习算法不要求类别集合固定不变，并且能够在新增类别训练样本集合达到一定数目时获得较好的分类效果。增量式学习是指在类别集合相同的前提下，当前增量步相对于上一个增量步，学习算法的分类性能有所提高，并在增量步达到一定规模时趋于稳定[7]。图3是实现适应新增类别与增量式学习的伪代码。

已知类别集合C={c_1, ..., c_m}，训练样本集为D={d_1,…, d_n}，其中，m为类别数，n为文档数。若当前分类样本d_x的系统推荐可信度最高的类别为c_r，用户判断后的类别是c_x，则有：

```
if(c_x ∉ C)   then
        D={d1,d2 ,…,dn ,dx};   n=n+1;
        C={c1,c2 ,…,cm ,cx};   m=m+1;
else if(cx!=cr)   then
        D={ d1,d2 ,…,dn ,dx };  n=n+1;
```

图3 实现适应新增类别与增量式学习的伪代码

4 实验与结果分析

4.1 实验设置

本文以航空领域为例，探讨领域本体构建中概念间关系的辅助判断问题。使用的数据源有：《中国航空百科词典》中除“综合类”以外的 8918 个术语释义项文档、航空领域专利摘要 24695 篇、2010 年维基百科中 21536 篇航空词条释义项。通过分析航空领域的特点，人工确定并标注分类所需初始概念关系类别如表 1：

表 1 初始关系类别及数目

初始关系类别	定义	举例	数量
部件关系	机械的一部分，由若干装配在一起的零件所组成	飞机——机翼	400 对
用途关系	表示航空产品所具有的功能	飞机——运输	400 对
材料关系	制造航空产品、零部件的物质	机翼——合金钢	400 对
属性关系	表示事物本身所固有的性质	飞机——机长	400 对
制造与工艺关系	将原材料加工成适用的产品的方法、过程和技术	黄铜——电铸法	400 对

从上表各类中抽取 30%作为测试概念关系集，70%作为训练概念关系集。分别采用传统 BM25 算法与改进的 BM25 算法对检索结果排序，其中 k 取 2.0，b 取 0.75。根据基于概念最短距离的分类样本提取方法获取概念关系集的测试样本，其中 n 取 5。使用 KNN 算法分别对两种排序结果中提取到的测试样本分类，其中相似度计算采用经典的 BM25 算法。

本文以 KNN 分类算法的准确率衡量改进 BM25 算法对检索结果的排序性能以及关系类别的推荐准确率，其中，以关系推荐中可信度值最高的关系类别作为最终类别，准确率定义如下：

$$\text{准确率} = \text{正确的分类结果} / \text{总的测试数据} \times 100\%$$

4.2 实验结果及分析

使用传统 BM25 算法与改进 BM25 算法排序后获得的样本在 k 取不同值时的实验结果如下：

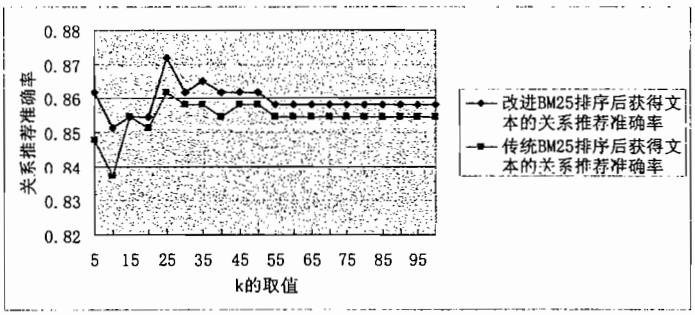

图 4 关系推荐准确率

从上图实验结果可以看出，在改进的 BM25 算法得到的排序结果中提取样本的 KNN 关系推荐准确率较传统的 BM25 算法高，在 k=25 时使用改进 BM25 算法排序后获取的分类样本的关系识别准确率达到 87.20%，而使用传统 BM25 算法的关系识别准确率为 86.15%，改进前后的分类结果均在 k=55 后趋于稳定。实验证明，改进的 BM25 算法能够有效的对检索结果排序，这不仅方便用户获得概念间关系的文本知识，也使得系统拥有较高的关系推荐准确率。

另外，为验证 KNN 算法对新增关系类别的适应能力，以及随关系推荐次数增加推荐效果越好的增量式学习能力，本文以部件关系类的训练语料为基础，按照用途、材料、属性、制造与工艺

的顺序依次增加训练语料类别，以步长 20 逐渐增加各类的训练语料数目，测试语料与训练语料的类别同步变化，图 5 是 k 取 25 时的实验结果：

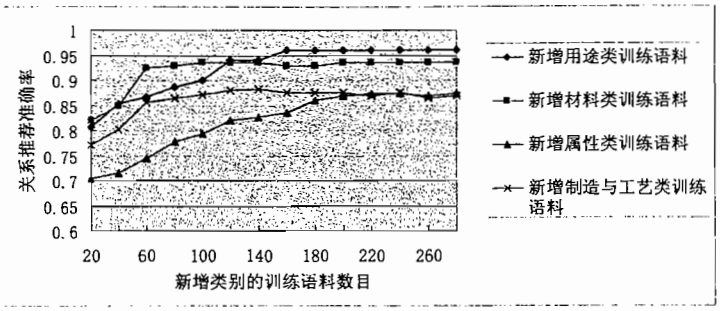

图 5　KNN 分类算法对于新增类别的适应能力

由上图可知，在类别集合不变的情况下，随着训练语料数目的增加，关系推荐准确率呈上升趋势，并在语料数目达到一定规模时趋于稳定。同时，对新增关系类别有较好的适应能力，虽然在新增关系类别时准确率存在一定程度的下降，但这与类别数目越多类间混淆越大的常理相符合。因此，使用本文方法能实现增量式的关系推荐，并且保证了对新增关系类别的适应性。

5　总结与展望

本文描述了一种领域本体构建中的关系辅助判断方法，该方法使用信息检索技术为用户提供关系判断的参考知识，通过改进 BM25 算法优化检索排序结果，使用基于概念最短距离的分类样本提取方法获取分类样本，并根据 KNN 算法推荐关系。实验结果表明该方法能够改进对信息检索结果的排序性能，从而为用户有效的提供关系参考以及拥有较高准确率的关系推荐，方便用户在构建本体过程中准确的判断概念间关系。

未来工作中，将考虑加入与概念所属类别相关的特征，保证本体中概念间关系构建的一致性；另外，考虑能否实现待判断关系的概念的启发式推荐，即与用户输入的两概念具有相关关系的概念推荐，并以此方法为基础构建领域本体关系库。

参 考 文 献

[1] Neches R. Enabling Technology for Knowledge Sharing[J]. AI Magazine, 1991.

[2] 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.

[3] Shamsfard M, Barforoush AA. Learning ontologies from natural language texts[J]. Int'l Journal Human Computer Studies, 2004, 60(1): 17-63.

[4] Missikoff M, Navigli R, Velardi P. Integrated approach for web ontology learning and engineering[J]. IEEE Computer, 2002, 35(11): 60-63.

[5] Maedche A, Staab S. The ontology extraction & maintenance environment Text-to-Onto[OL]. In: Proc. of the ICDM 2001 Workshop on the Integration of Data Mining and KnowledgeManagement. 2001. http://cui.unige.ch/~hilario/icdm-01/DM-KM-Final/Volz.pdf.

[6] 杜小勇, 李曼, 王珊. 本体学习研究综述. 软件学报[J], 2006, 17(9): 1837-1847.

[7] 刘浩公, 蔡东风, 张桂平. 一种基于词典的领域本体建立方法[J]. 通讯和计算机, 2007(4): 61-64.

[8] 蔡丽宏, 马静, 吴一占. 基于 OWL 的本体半自动进化研究[J]. 情报学报, 2011, 30(1): 56-60.

[9] Wiefinga B J,Schreiber A T, Wielemaker J,et al. From thesaurus to ontology[C]. Proceedings of the 1st International Conference on Knowledge Capture. [2008-11-20]. http://www.cs.vu.nl/guus/papers/Wielinga01a.pdf.

[10] HEARST M A. WordNet:an electronic lexical database[M]. Cambridge: MIT Press, 1998.

[11] Maedche A, Staab S. Discovering conceptual relations from text. In: Horn W, ed. Proc. of the ECAI 2000[J]. Amsterdam: IOS Press, 2000: 321-325.

[12] Kavalec M, SvateK V. A study on automated relation labeling in ontology learning[J]. In: Buitelaar P, Cimiano P, Magnini B, eds. Ontology Learning from Text: Methods,Evaluation and Applications. Amsterdam: IOS Press, 2005.

[13] Kavalec M, SvateK V. A study on automated relation labeling in ontology learning[J]. In: Buitelaar P, Cimiano P, Magnini B, eds. Ontology Learning from Text: Methods,Evaluation and Applications. Amsterdam: IOS Press, 2005.

[14] KHAN L,LUO F. Ontology construction for information selection[C]. Washington D C; Proceedings of 14th IEEE International Conference on Tools with Artificial Intelligence, 2002.

[15] Robertson S. E., Walker S., Beaulieu M. Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive track[A]. Proceedings of the 7th Text Retrieval Conference[C]. Gaithersburg, 1998: 253-264.

[16] Cover T, Hart P. Nearest neighbor pattern classification[J]. IEEE Trans. Inform. Theory IT-11: 21-27, 1967.

[17] 罗长升, 段建国, 郭莉. 基于推拉策略的文本分类增量学习研究[J]. 中文信息学报, 2008, 22(1): 37-43.