情报工程 ISSN: 2095-915X
第1卷 第6期 74-82 2015年12月

TECHNOLOGY INTELLIGENCE ENGINEERING
ISSN: 2095-915X Vol.1 No.6 74-82 Dec 2015

doi:10.3772/j.issn.2095-915x.2015.06.011

利用多策略模糊综合评判的术语关系识别方法研究

殷希红，乔晓东，张运良

（中国科学技术信息研究所 北京 100038）

摘要：利用多策略模糊综合评判的方法进行术语关系识别，首先采用多种相似度计算方法计算术语的相似度，然后利用连续属性离散化方法确定关系类别及阈值区间的划分，利用样本分布概率确定区间对类别的隶属度，利用粒子群算法和交叉验证法确定因素权重，最后利用模糊综合评判方法将所有相似度计算方法的计算结果进行融合处理，实现术语关系的识别。本研究将以中国科学技术信息研究所已有的新能源汽车领域汉语科技词系统中的术语作为测试集，用准确率、召回率和F值对关系识别的结果进行评价，论证该方法的有效性。

关键字：模糊综合评判，多策略，关系识别，相似度

分类号：G254 TP391

Research on the Method of Term Relationship Recognition Using Multi-Strategy Fuzzy Comprehensive Evaluation

Yin Xihong, Qiao Xiaodong, Zhang Yunliang

(Institute of Scientific and Technical Information of China, Beijing 100038,China)

Abstract: The multi-strategy fuzzy comprehensive evaluation is used to recognize the term relations. First, a variety of similarity calculation methods are used to calculate the terms similarity. Then, relations and threshold interval are identified by discretization algorithm of continuous attributes, the sample distribution probability

基金项目：国家自然科学基金项目“面向特定情报分析应用的知识组织系统快速构建关键问题研究”（71203208）
国家科技支撑计划项目“面向科技情报分析的信息服务资源开发与支撑技术研究”(2015BAH25F01)

is used to determine the membership degree of the interval to the relations, and the weight of elements are determined by particle swarm algorithm and cross validation method. All the calculation results are composited by fuzzy comprehensive evaluation method to recognize the term relations. At last, the precision, recall and F value are used to evaluate the effect of the results. This experiment regards the Chinese scientific and technical vocabulary system (new energy vehicles) as the test data. The result shows that the method can recognize the term relations effectively.

Key words: Fuzzy comprehensive evaluation, multi-strategy, term, relation recognition, similarity

1 引言

在词系统构建时，都需要对词汇关系进行分类，将词汇划分到对应的关系类别下，因此词汇关系识别有重要意义。但目前，相似度计算方法侧重于对术语间关系的简单判断，很少对术语的相似关系做深度细分。术语的相似关系是术语间的同义关系、反义关系、缩写和全称关系、上下位关系等特定关系的统称。术语相似关系识别方法大致可分为三种：人工识别、半自动识别和自动识别。人工识别方法虽然可行，准确率高，但是需要手工构建，耗时费力。半自动识别和自动识别虽然构建周期短，效率高，但是仍然存在以下不足：第一，仅能针对单一关系或者少量关系识别，如同义关系识别或者上下位关系识别[1]，识别关系数量有限，而词汇关系不局限于这几种关系，还包括组成关系、控制关系、时间关系、空间关系等。第二，识别关系笼统，未细化，构建词系统时还需要进一步加工处理，如相似关系识别，即两个词汇只要具有某种关系便为相似关系，而不能确定具体属于何种相似关系。第三，相似关系识别方法大部分只利用一种相似度计算方法，识别出的结果往往具有片面性。

因此本文正是基于中国科学技术信息研究所已有的新能源汽车领域汉语科技词系统中的术语及术语定义信息，提出采用多策略相似度计算方法，并结合模糊综合评判方法将相似度计算结果进行融合处理，然后利用模糊综合评判方法的计算特征实现术语关系的细分。

2 国内外多策略关系识别研究现状

多策略词汇关系识别方法能够融合几种算法的优势，避免单一算法的不足，提高关系识别结果的准确率和召回率。通过对国内外多策略词汇关系识别的调查研究，将采用多策略进行词汇关系识别的方法大致归纳为以下三种：

（1）整合去重

陆勇等[2]综合利用字面相似度算法、特征模式匹配算法和 PageRank 链接分析算法对百度百科语料库资源进行同义关系自动获取，然后将各方法的同义关系识别结果整合去重，实验表明三种方法同义关系识别结果差别较大，重合率低，可以有效地实现同义关系互补识别。但是整合去重的融合方法，虽然能够提高同义关系识别的召回率，但是会影响准确率。

（2）加权融合

马海昌等[3]以搜狗实验室网站提供的关于新闻的语料作为研究对象，结合潜在语义分析

和点互信息的方法进行同义词抽取，将词汇在LSA分解结果的相似度和互信息加权相加融合起来，当结果大于某阈值时，则认为两者为同义关系。实验证明结合方法较单一方法在准确率、召回率和F值均有所提高。后来，马海昌等[4]又将词汇字面相似方法和PageRank链接方法采用加权相加融合的方式从百度百科经济领域中获取同义词，实验表明该融合方法可从大规模语料中获取大量同义词集合，较单一方法在同义词识别的准确率、召回率和F值上均有所提高。Henriksson等[5]为克服单一模型仅适用于在特定领域中抽取同义词的不足，将不同的分布模型进行合并，然后分别应用于单语语料库和双语语料库，采用余弦值相加的融合策略进行同义词识别，实验证明组合模型优于单一模型。采用多策略加权融合的方式虽然能提高关系识别的效果，但是如果权值设置不合理仍然会影响到融合效果。

（3）启发式学习

Neshati等[6]人采用基于WordNet的相似度测度、基于窗口共现的相似度测度、基于句法共现的相似度测度及基于web的互信息测度，将针对一对语词各方法计算出来的相似度表达成一个4维向量，然后使用旅游分类系统作为训练集，通过路径相似度算法计算在分类系统中词对之间的相似性，并作为神经网络模型学习评判各种相似度计算测度权重的依据。然后使用金融领域的分类系统作为测试集，对于每一对语词产生的相似度向量，利用神经网络学习阶段产生的权重输出一个复合的相似度测度。Bollegala等[7]提出了基于Web搜索引擎的术语相似度计算，分别使用Jaccard，Overlap，Dice和PMI四个相似度指标来衡量语对在搜索引擎命中数方法的相似度，使用n-Gram的模板匹配来衡量语词对在摘要方面的相似度。选择词对共现频率最高的200个模板与四个相似度计算指标构成204维向量作为一个二类SVM分类器的输入，产生最终的相似度判断结果。而二类SVM分类器由从WordNet中抽取的被认为相似和不相似的语词对构成的204维向量集合作为训练语料进行训练得到。

从以上可以看出，多策略词汇关系识别比单一方法有较大的优势，多策略的词汇关系识别可以以互补的方式组合几种计算方法，融合各个算法的优势，提高词汇关系识别的效果。但是国内外多策略的词汇关系识别研究基本上只针对同义关系识别，而对其他关系识别和关系细分的研究尚有不足。因此，为进一步研究词汇关系识别，本文提出基于模糊综合评判的多策略关系识别方法。

3 多策略模糊综合评判关系识别方法

3.1 模糊综合评判原理

模糊综合评判法[8,9]是一种基于模糊数学的综合评判方法。该综合评判法根据模糊数学的隶属度理论把定性评价转化为定量评价，即用模糊数学对受到多种因素制约的事物或对象做出一个总体的评价。首先确定由多种因素组成的模糊集合（因素集U），再设定这些因素所能选取的评审等级，组成评语的评判集合（评判集V），分别求出各单一因素对各个评审等级的归属程度（模糊矩阵），然后根据各个因素在评价目标中的权重分配，通过计算（模糊矩阵合成），求出评价的定量解值，上述过程即为模糊综合评判。

3.2 多策略模糊综合评判方法设计

模糊综合评判方法依赖于模糊数学的隶属

度理论，而隶属度的确定通常采用综合加权法、对比排序法、指派法、模糊统计法等。隶属函数的确定过程本质上应是客观的，但是这些方法均存在着一定的人工干预，因此为了增加确定隶属度函数的自动化程度，本文引入连续属性离散化方法 [10,11] 将各个因素的取值范围分别划分成不同的区间，得到在不同区间内样本的分布情况，以及在同一区间内样本对应不同关系类别的分布情况，然后计算样本的分布概率，并将其作为各个因素在不同区间下对应评判集的隶属度，概率越大表明隶属度越大。评判因素的权重设置通常采用德尔非法、层次分析法等，但是这些方法存在着一定的主观性，不同的权重设置会影响到模糊综合评判的最终结果，因此为增加权重设置的客观性，本文采用粒子群算法 [12] 和交叉验证方法对训练集进行训练，通过粒子寻优的方式，得到各个因素的最优权重分配。

多策略模糊综合评判方法的计算流程如图 1 所示：

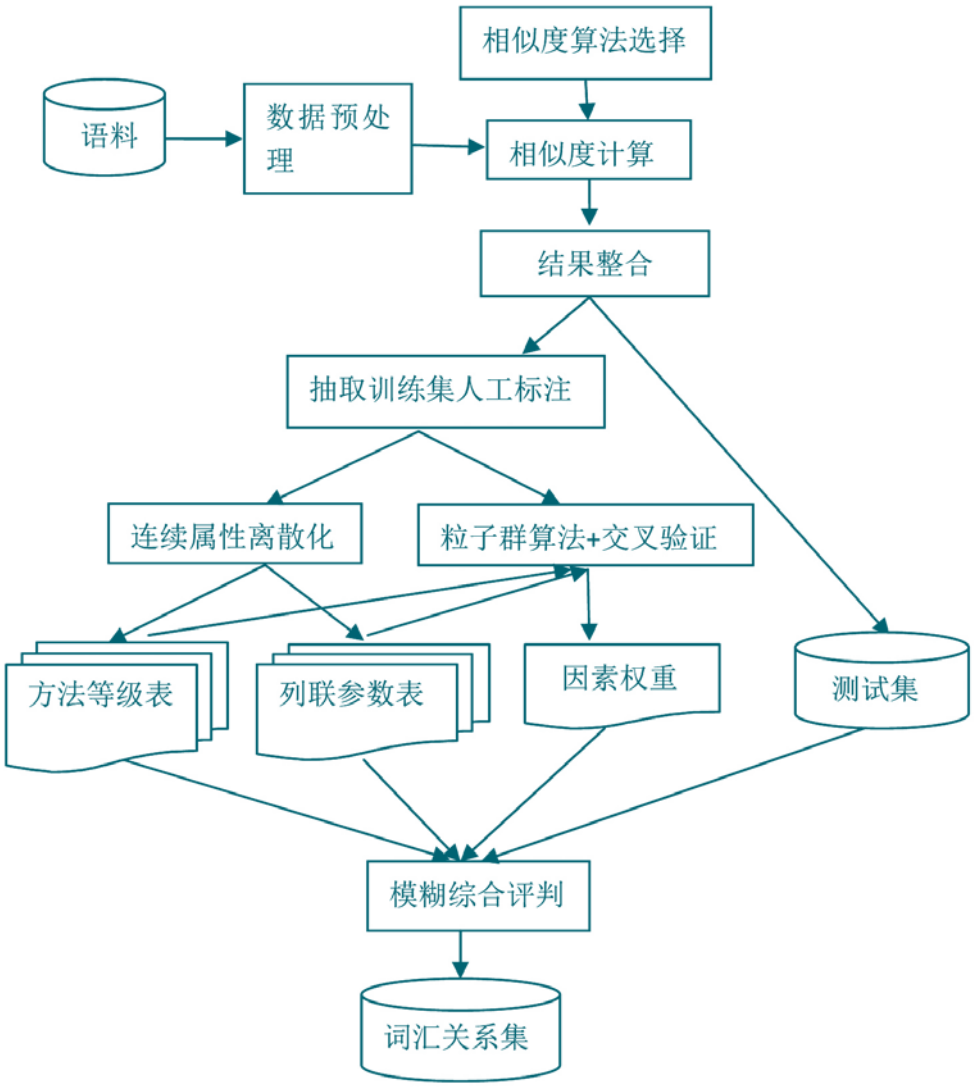

图 1 多策略模糊综合评判方法计算流程图

其计算步骤如下：

步骤 1：确定因素集 U={u_1，u_2，…，u_m}，以采用的各个相似度算法作为评价因素，相似度算法可根据处理数据和计算目的自行选择，其中：

u_i（i=1,2,…，m），表示第 i 种相似度算法；

步骤 2：确定评判集 V={v_1，v_2，…，u_m}，根据各个相似度算法计算结果，确定评判等级为 n 个等级，每个等级对应为词汇间的一种关系，等级划分粒度根据关系的细分程度进行确定，其中：

v_i（j=1,2,…,n），表示词汇间的第 j 类关系；

步骤 3：确定评价因素权重

确定评价因素的权重即为确定多策略融合方案中各个相似度算法的权重。在此，使用交叉验证的方法首先将训练集等分为十份，每次取其中一份作为验证集，其他九份作为训练集，采用粒子群算法，将因素权重作为求解参数进行初始化，以求平均准确率作为目标函数，粒子群算法通过寻求目标函数的最大值得到达到最高的平均准确率所对应的一组参数值，即为各个相似度计算方法的一组最优权重，作为最终各因素的权重。

步骤 4：建立模糊关系矩阵

a、首先从要处理的数据中随机抽取部分数据作为训练集，分别利用每个相似度计算方法进行计算，针对一对词汇均会得到一组相似度值为：

sim($term_1$, $term_2$)={sim_1, sim_2,…,sim_m }

b、人工确定训练集中各个方法计算的相似度值所对应词汇间的关系。

c、根据训练集得到的计算结果利用连续属性离散化方法进行阈值区间划分，由于相似度计算方法的值域为 [0,1]，所以 e_0=0，e_1=1。连续属性离散化方法是利用训练集计算出的结果，求出一组最优的（e_0，e_1，…，e_L）值，作为划分区间的边界，并且由此得到在不同阈值区间中训练集在关系类别中的分布情况，如表 1 所示：

表 1 阈值关系列联表

		因素 u_i 的区间划分					合计
		$[e_0, e_1)$	…	$[e_{r-1}, e_r)$	…	$[e_{L_i-1}, e_{L_i}]$	
关系类别	v_1	q_{11}	…	q_{1r}	…	q_{1L_i}	q_{1+}
	⋮		⋮		⋮		⋮
	v_j	q_{j1}	…	q_{jr}	…	q_{jL_i}	q_{j+}
	⋮		⋮		⋮		⋮
	v_n	q_{n1}	…	q_{nr}	…	q_{nL_i}	q_{n+}
合计		q_{+1}	…	q_{+r}	…	q_{+L_i}	M

表 1 中第 i 个相似度计算方法的值域被分为 L_i 个区间，其中 q_w 为在训练集中相似度值的取值为 [e_{r-1}, e_r)，关系类别为 的所有样本点个数。

d、根据训练集在列联表中的分布情况，以在同一阈值区间内对应的不同类别的样本点分布的概率作为该阈值区间对应不同关系类别的隶属度，其计算如公式 1 所示：

$$d_{jr} = \frac{q_{jr}}{q_{+r}} \text{ (j=1, 2, . . n)} \qquad (1)$$

公式（1）表示当相似度计算方法的取值在第 r 个阈值区间，即 [e_{r-1}, e_r) 时，对于不同关系的隶属度。

步骤 5：综合评判

a、假定采用各个相似度算法得到的相似度组成因素集合为 u=(u_1, u_2 ,…, u_m)，首先查找各个因素输出值对应的阈值区间分别为 r_1, r_2 ,…, r_m。

b、查找各个阈值区间对应不同关系的隶属度，然后构成模糊关系矩阵为：

$$R=\begin{pmatrix} d_{1r_1} & d_{2r_1} & \cdots & d_{nr_1} \\ d_{1r_2} & d_{2r_2} & \cdots & d_{nr_2} \\ \cdots & \cdots & \cdots & \cdots \\ d_{1r_m} & d_{2r_m} & \cdots & d_{nr_m} \end{pmatrix}$$

c、以权重分配 W 与模糊关系矩阵构成模糊评判模型：

$$B=W\circ R=(w_1, w_2, \cdots, w_m)\circ \begin{pmatrix} d_{1r_1} & d_{2r_1} & \cdots & d_{nr_1} \\ d_{1r_2} & d_{2r_2} & \cdots & d_{nr_2} \\ \cdots & \cdots & \cdots & \cdots \\ d_{1r_m} & d_{2r_m} & \cdots & d_{nr_m} \end{pmatrix} = (b_1, b_2, \cdots, b_m)$$

d、B 中最大值对应的关系即为该多策略相似度计算方法得到的关系。

4 实验及结果分析

4.1 数据来源及预处理

选取“新能源汽车领域”汉语科技词系统 [13] 中的术语及其定义作为语料。该书收录中文核心词 6117 条，在此基础上，经过增补，截至 2013 年 1 月，核心词数量为 6216 条，其中含有定义的不重复术语为 5331 条，因为一条术语可能包含两条以上的定义，术语定义总共为 5888 条。对术语及定义做去噪及分词处理，保存到数据库中备用。

4.2 相似度算法实验

分别从字面、句法、语义三个角度选取字面相似度算法，模式匹配算法和定义匹配算法计算术语对的相似度值。其中字面相似度算法采用侯汉清 [14] 提出的经典的加权相似度计算方法；而模式匹配算法则借鉴陆勇、侯汉清 [15] 的模式匹配识别汉语同义词的思想，对语料中术语的定义方式进行归纳总结，设计特征模式，编写术语关系提取程序；定义匹配算法采用殷希红等 [16] 提出基于定义的相似度计算方法。然后将三种方法的相似度计算结果进行整合处理，得到针对每一术语对的一组相似度值。如表 2 所示：

表 2 三种相似度计算方法结果整合表

term1	term2	sim1	sim2	sim3
M-ETBE 汽油	ETBE 汽油	0.35	0.81	0
安全性	被动安全性	0	0.7	0.7
安全性	生态安全性	0	0.7	0.7
安全性	主动安全性	0	0.7	0.7
变速器	齿轮变速器	0.83	0.7	0
变阻器	频敏变阻器	0	0.7	0.7
采样点	固定采样点	0	0.7	0.7
采样点	流动采样点	0.38	0.7	0
测距器	联动测距器	0	0.7	0.7
车架	正车架	0.78	0.74	0
车桥	后车桥	0.4	0.74	0

4.3 多策略模糊综合评判实验

模糊综合评判中因素集 U={sim1，sim2，sim3}，sim1 为基于定义匹配算法得到的相似度值，sim2 为基于字面相似度算法得到的相似度值，sim3 为基于模式匹配算法得到的相似度值。由于确定要识别的词汇关系分别为同义关系、层级关系、同类关系、无关系四种，因此评判集 V={v_1, v_2, v_3, v_4}。首先从整合后的数据中随机抽取 500 条作为训练集，200 条作为测试集，并根据要识别的关系对术语对的关系进行人工标注（见表 3），然后利用多策略模糊综合评判方法对训练集进行训练，得到针对单一相似度计算方法的因素等级表、列联参数表及各个因素的权重，最后输入测试集，采用模糊综合评判方法，得到各个术语对所对应的关系（见表 4）。

表 3 中 0 表示层级关系，1 表示同类关系，2 表示同义关系，3 表示无关系。term1 和 term2 为一个术语对，sim1，sim2，sim3 为三种方法计算得到的相似度值。

表 3 术语关系标注结果

relation-id	term1	term2	sim1	sim2	sim3
1	CNG 管路	LPG 管路	0.60	0.00	0.00
2	主动悬挂	主动悬架	0.65	0.00	0.00
2	DSC 系统	动态稳定性控制系统	1.00	0.00	0.00
0	海底电缆	海底电力电缆	0.00	0.00	0.70
2	MPV 车型	多功能汽车	0.89	0.00	0.00
1	N 型半导体	P 型半导体	0.71	0.85	0.00
1	单蛙绕组	双蛙绕组	0.67	0.00	0.00
0	背场背反射太阳电池	背反射太阳电池	0.00	0.91	0.00
3	倍率放电	脉冲充电	0.33	0.00	0.00
3	电容起动电动机	初级绕组	0.33	0.00	0.00

表 4 多策略模糊综合评判关系识别结果

relation_id	term1	term2
3	四冲程	二冲程柴油机
1	不等速万向节	准等速万向节
1	常规型电动汽车	电动轮型电动汽车
0	自整角接收机	差动自整角接收机
1	动力转向系统	电动助力动力转向系统
2	气体继电器	瓦斯继电器
2	对位芳酰胺纤维	代石棉垫片
0	消石灰	熟石灰
1	干荷蓄电池	干式荷电蓄电池

4.4 结果分析

用准确率和召回率对多策略模糊综合评判关系识别结果和基于 SVM 的关系识别结果进行评估，准确率是指识别的正确的术语关系占识别的总的该关系的比例；召回率是指识别的正确的术语关系占测试集中存在的该关系的比例。

$$Precision=\frac{correctly\ found\ relation<term1,term2>}{total\ found\ relation<term1,term2>}$$

$$Recall=\frac{correctly\ found\ relation<term1,term2>}{total\ relation<term1,term2>}$$

$$F1\text{-}measure=F1=\frac{2*Precision*Recall}{Precision+Recall}$$

表 5 多策略模糊综合评判术语关系识别效果

关系类型	识别数量	正确数量	测试集中数量	precision	recall	F1
层级关系	50	32	61	64.0%	52.5%	0.58
同类关系	85	43	61	50.6%	70.5%	0.59
同义关系	38	34	49	89.5%	69.4%	0.78
无关系	27	17	29	63.0%	58.6%	0.61
总计	200	126	200	63.0%	63.0%	0.63

从表 5 中可以看出该多策略模糊综合评判的关系识别效果较好，对于各种关系的识别均可达到较高的准确率和召回率，其中对于同义关系识别的准确率最高为 89.5%，同类关系识别的准确率最低为 50.6%，同类关系的识别的召回率最高为 70.5%，层级关系识别的召回率最低为 52.5%。从 F 值可以看出本方法对于同义关系的识别效果最好。

5 结语

基于模糊综合评判方法提出利用多策略相似度算法融合进行术语关系识别的方法，该多策略关系识别方法可以弥补单一相似度计算方法的不足，融合多个算法的优势。并且可以达到较高的准确率、召回率，说明该方法有效可行。但是由于本方法会受所选择的相似度计算方法及相似度计算方法数量的影响，因此后续将加入更多的相似度计算方法进行实验或者选择几种较好的方法进行融合。另外，该方法不仅适用于术语关系识别，对于其他分类问题同样适用，仅需要改变对应的因素集及评判等级，因此该研究也为后续分类问题的研究提供思路。

参考文献：

[1] 张巍，于洋，游宏梁．面向词汇知识库自动构建的概念术语关系识别 [J]. 现代图书情报技术，2009（11）：10-16.

[2] 陆勇，章成志，侯汉清．基于百科资源的多策略中文同义词自动抽取研究 [J]. 中国图书馆学报，2010,36(185):56-62.

[3] 马海昌，张志昌，赵学锋，等．结合潜在语义分析与点互信息的同义词抽取 [J]. 电脑知识与技术，2014,10(1):128-132.

[4] 马海昌，张志昌，赵学锋，等．面向经济领域的同义词获取融合方法研究 [J]. 科学技术与工程，2014,14(15):207-211.

[5] Henriksson A, Hans Moen, Skeppstedt M, et al. Synonym extraction and abbreviation expansion with ensembles of semantic spaces [J]. Journal of Biomedical Semantics, 2014,5(6):1-25.

[6] Neshati M, Hassanabadi LS. Taxonomy

Committee, 2007:757-766.

[8] 李洪兴，汪培庄. 模糊数学 [M] . 北京：国防工业出版社 , 1994.

[9] 谢季坚 , 刘承平. 模糊数学方法及其应用[M]. 武汉: 华中科技大学出版社 , 2005:31-36.

[10] 陈秉正 , 韩春鹏 . 归纳式学习中连续型数据的区间划分问题 [J]. 系统工程理论与实践，2001 (4) : 1-7.

[11] Chan K C C, Wong A K C. Class-dependent discretization for inductive learning from continuous and mixed-mode data[J]. IEEE Trans. Pattern Analysis and Machine Intelligence, 1995,17(7):641-651.

[12] Kennedy, J., Eberhart, R.. Particle Swarm Optimization[C]//Proceedings of IEEE International Conference on Neural Networks IV,1995:1942-1948.

[13] 贺德方 , 乔晓东 , 朱礼军 , 等 . 汉语科技词系统 (新能源汽车卷) [M]. 北京 : 科学技术文献出版社 , 2012.

[14] 侯汉清 , 吴志强 . 利用字面相似度识别汉语同义词的实验 [C]// 信息服务业的信息化 , 2001:222-229.

[15] 陆勇 , 侯汉清 . 基于模式匹配的汉语同义词自动识别 [J]. 情报学报 ,2006, 25(6): 720-724.

[16] 殷希红 , 乔晓东 , 张运良 . 利用术语定义的汉语同义词发现 [J]. 现代图书情报技术 , 2014,30(4): 41 - 47.