基于交通领域知识网络的词汇语义相似度计算①

黄 浩, 陈怀新

(中国电子科技集团第十研究所, 成都 610036)

摘 要: 针对传统基于 wordnet 的词汇语义相似度计算方法中隔离抽象词汇和具象词汇, 以及片面依赖上下义关系的不足, 提出了基于交通领域知识网络的词汇语义相似度计算方法. 基于上下义、工具-工具对象、部件-整体等概念关系准则构建了交通词汇的知识网络图谱, 提出了修正的平均路径长度参量计算网络中词汇的语义相似度, 得到更高的语义一致性结果. 实验表明, 在Finkelstein的353对词汇集上, 本文算法能够获得比传统方法更符合人工判断的语义相似度.

关键词: 词汇语义相似度; 领域知识网络; 平均路径长度; wordnet; 概念关系准则

Measuring Semantic Similarity of Words Based on Traffic Field Knowledge Network

HUANG Hao, CHEN Huai-Xin

(China Electronics Technology Group Corporation No.10 Research Institute, Chengdu 610036, China)

Abstract: The traditional way of calculating word semantic similarity is based on wordnet structure, which has a huge gap between physical concept and abstract concept, and only considering concepts' hyponymy. To solve the problem, a novel word similarity calculation algorithm based on traffic field words relation network is proposed in the paper. 10 kinds of concept relationships, including concepts of hyponymy, tool-tool object relationship, standard parts-overall and so on, are used to build traffic words knowledge network. Then modified average path length parameter is used to calculate words' semantic similarity, which accords with people's judgement. The experiment based on Finkelstein's 353 word pairs shows that the algorithm achieves more accurate word semantic similarity.

Key words: word semantic similarity; field knowledge network; average path length; wordnet; concept relationship rule

随着语义相似度是两个语言对象在各种语言互动类型下的语义互动强度[1]. 词汇作为自然语言最基本的单位, 它们之间的语义相似度计算是一项十分重要的基础工作, 在信息检索、机器翻译、图像标签排序和图像标签推荐等领域都有着广泛的应用.

语义相似度的计算共有两类方法: 基于分布相似性统计的相似度计算和基于知识资源结构分析的相似度计算. 前者基于这样一个假设: 相似的词汇出现在相似的上下文中. 统计词汇对在语料库文本窗口(通常为句子、段落或者篇章)中共同出现的频次, 频次越大, 相似度越大. 孙叔琦[2]和Mohammad[3]分别采用共生关系和平均互信息方法来计算词汇对的语义相似度. 但是, 这种方法计算结果的准确性受到语料库规模和所选计算公式的影响.

基于知识资源结构分析的相似度计算方法通过分析专家知识库组织结构的规律, 提出合理计算公式来量化知识库中词汇的相似关系. wordnet 是其中应用最为广泛的知识库, 由普林斯顿的语言学家和心理学家编撰, 涵盖了近 117000 的英文词汇. wordnet 以义项为单元, 通过上下义和整体部分关系连接所有义项, 构成了一个有层次结构的词汇网络. 其中, 上下义关系占比 90%以上, 生成了以"entity(事物)"为根节点的大型树结构. 目前, 绝大多数相似度计算方法都是基于树得到的, 常见的有基于义项间路径长度的方法[4], 基于最深公共父节点信息内容的方法[5]和基于义项释义重合度[6]的方法等. 但是, 在实际应用中, 许多词汇

① 收稿时间:2016-06-21;收到修改稿时间:2016-08-08 [doi:10.15888/j.cnki.csa.005652]

的计算结果并不符合人的语义判断, 存在以下不足: 第一, wordnet 分为“抽象事物”和“具象事物”两个子树, 造成了抽象概念和具体事物的天然隔离, 使“交通”-“汽车”的相似度远小于“交通”-“亚洲”的相似度; 第二, 即使在具象名词分支, “汽车”-“轮子”、“公路”-“汽车”等关系紧密的词汇也因为单一的上义下义关系而变得相似度很低.

针对以上不足, 本文提出了基于交通领域知识网络的词汇语义相似度计算方法. 该方法首先搜集某一领域的常用词汇, 通过上下义、工具-工具对象、场所-事件等 10 种关系准则多角度表达领域知识潜在联系, 然后基于词汇在关系网络中的路径长度计算它们的语义相似度, 使计算结果更符合人的语义判断.

1　基于wordnet的词汇语义相似度

Wordnet 是以上下义关系为主的分类关系树. 传统的方法多基于义项在树中的结构关系来计算语义相似度, 共分为 2 类: 基于路径长度的算法和基于公共父节点信息内容的算法.

1.1 基于路径长度的义项语义相似度

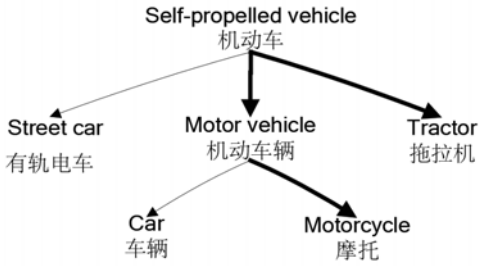


图 1　Wordnet 分类树中部分义项的组织结构

路径长度指的是两个义项在分类树中形成的一条通路上所包含边的个数. 在图 1 中, “摩托”和“拖拉机”的一条通路由黑色加粗的线段表示, 路径长度为 3. Hirst[7]指出义项在分类树中的最短路径越短, 相似度越大, 并直接利用路径长度计算相似度, 公式如下:

$$\text{snetsim}(s_1,s_2) = -\log(len(s_1,s_2)) \tag{1}$$

其中, snetsim(s_1,s_2)表示义项 s_1 和 s_2 的语义相似度, len(s_1,s_2)表示和 s_2 的路径长度.

Yu[8]认为相似度的大小不仅与路径长度相联系, 还与该节点在分类树中的深度有关. 在相等路径长度的条件下, 义项的深度越大, 概念越具体, 它们之间的区别也越小, 语义相似度需要加强. 计算公式如下:

$$\text{snetsim}(s_1,s_2) = -\log\frac{dep(s)}{len(s,s_1)+len(s,s_2)-2*dep(s)} \tag{2}$$

其中, s 为义项 s_1 和 s_2 的最深公共父节点, $dep(s)$表示义项 s 的深度.

Leacock 和 chodorow[9]则以 wordnet 中最大的深度作参考, 提出如下计算公式:

$$\text{snetsim}(s_1,s_2) = -\log\frac{len(s_1,s_2)}{2*D} \tag{3}$$

其中, D 表示 Wordnet 分类树的全局最深节点的深度.

1.2 基于信息内容的义项语义相似度

借鉴信息论中信息熵的概念, 基于信息内容(Information Content, IC)的算法将两个义项的最深公共父节点(Least Common Ancestors, LCA)所包含的信息量作为两者之间的语义相似度. 计算公式如下:

$$IC(s) = -\log(Pr(s)+\sum_i Pr(s_i)) \tag{4}$$

其中, s 为义项的 LCA, $IC(s)$表示 s 的信息熵, $Pr(s)$表示 s 在语料库中出现的频率, s_i 表示 s 的子节点. 义项的信息内容由它本身以及所有的子节点概率和表示. LCA 的深度越浅, 包含的子节点越多, 计算结果越小. 这正符合 wordnet 树结构的特点, 树中每一层节点都是对下一层子节点概念的抽象. 概念越抽象, 所含的信息量越小. Lin[10]直接使用 LCA 的信息量作为相似度的大小. Formica[11]在计算中加入了节点各自的信息内容, 公式如下:

$$\text{snetsim}(s_1,s_2) = \frac{2*IC(s)}{IC(s_1)+IC(s_2)} \tag{5}$$

Jiang 的计算方式避免了结果中过多的小值:

$$\text{snetsim}(s_1,s_2) = \frac{1}{IC(s_1)+IC(s_2)-2*IC(s)} \tag{6}$$

一般地, 词语由多个义项构成, 如何从义项的语义相似度得到词汇的相似度, 常用的方法是取所有义项组合中语义相似度的最大值作为词汇的语义相似度. 该方法计算简单, 在很多应用中也符合词义模糊处理的需要. 假设词汇 w_1 的义项为 $s_i(0<i\le m)$, w_2 的义项为 $s_j(0<j\le n)$, w_1 和 w_2 的词汇语义相似度计算公式如下:

$$\text{wordsim}(w_1,w_2) = \max(snetsim(s_i,s_j)) \tag{7}$$

2　本文算法

传统的方法虽然在相似度计算上取得了一定的效果, 但也存在着很多的问题:

1) 基于信息内容的算法同时需要专家知识库和语料库的支持, 加大了计算开支. 而且, 词汇信息量的计算严重依赖于语料库的质量. 不同的语料库所计算的结果可能相差很大.

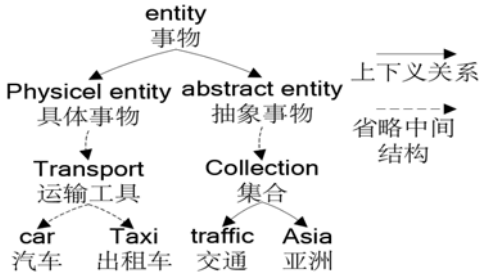


图 2 Wordnet 中具体事物与抽象事物的部分结构

2) 如图 2 所示, 对于抽象词汇和具象词汇的相似度计算, 无论是基于路径长度还是基于信息内容的方法, 都存在天然的“弱相似性”, 这在许多场合下并不合理.

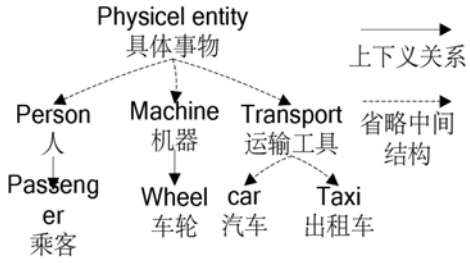


图 3 Wordnet 中具体事物分支部分结构

3) 如图 3 所示, 即使都为具象词汇, 片面依赖分类学的关系, 许多在内涵上有很强关联性的词汇(汽车-车轮、出租车-乘客等)也无法获得符合人工判断的相似度.

4) wordnet 中的词汇虽然覆盖面广, 但在某个领域内并不详尽, 很多术语不在其列, 而且词汇分布发散, 相互关系不易管理.

针对以上不足, 构造了交通领域的知识关系网络用以计算词汇的语义相似度. 改进如下:

1) 采用不依赖语料库的基于路径长度的算法;

2) 不再区分抽象和具象事物的词汇, 而是采用包含着语义信息的“情景发生”方式重新组织词汇网络;

3) 在上下义关系的基础上, 增加了部件-整体、属性-宿主等 9 种关系, 从不同角度还原人工语义判断的依据;

4) 搜集单个领域内完善的专业知识, 按照上述方法构建词汇网络, 并推广至其它的领域.

2.1 交通词汇知识网络的构建

为了打破抽象词汇和具象词汇的天然壁垒, 不再按照抽象事物和具体事物的标准划分词汇, 而是根据“情景发生”的方式组织词汇网络. 本文模拟事件发生的三要素(对象, 行为和环境), 将常用的交通领域词汇按照交通主体(包括人和交通工具)、交通行为和交通环境(交通地点、交通发生时间等)划分.

董振强在编撰知网[12]时, 曾指出词汇之间除了简单的分类学关系外, 还有部件-整体关系、属性-宿主关系、关系主体-事件关系、场所-事件关系、时间-事件关系、值-属性、实体-值和工具-工具对象关系等. 这些关系反映了我们感知词汇语义关系的不同侧面. 比如“汽车”和“驾驶”是一种关系主体-事件的关系, 但在 wordnet 中, 它们分别属于具体事物和抽象事物, 相关性很低. “赛车”和“快”在wordnet中分属于不同的词性树, 无法进行相似度计算, 而实体-值的关系体现了它们语义相关的一面. 交通领域词汇在这 10 种关系的实例如表 1 所示. 值得指出的是, 这些关系概括了人们语义判断时一般性的词汇关联模式, 不止适用于交通运输领域, 借助相应的专业知识, 可以方便地推广至其它领域.

表 1 10 种基本语义关系以及交通词汇实例

关系名称	实例
上下义	运输工具-汽车
部件-整体	车轮-有轮运输工具, 挡风板-汽车
工具-工具对象	汽车-司机, 赛车-赛车手
关系主体-事件	行人-闯红灯, 汽车-堵车
受事-事件	自行车-骑行
场所-事件	车站-等车/上下车
时间-事件	高峰时间-堵车, 工作日-通勤
属性-宿主	汽车-速度, 红灯-颜色
值-属性	红色-颜色, 快-速度
实体-值	红灯-红色, 汽车-快

交通领域知识网络的构造分为两步. 首先利用上义和下义关系将交通领域的词汇组织为网络的基本骨架, 然后依次考察每对词汇之间是否存在其它的联系, 如果存在, 则在词汇对之间添加新的关系连线. 图 4 展示了本文构建的部分词汇的关系网络图.

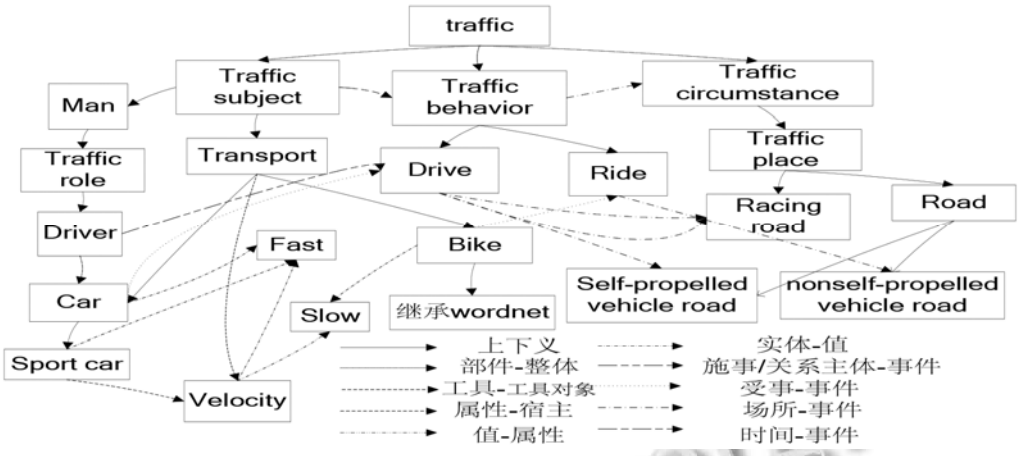


图 4　交通领域关系网络图部分结构

相比传统的 wordnet 分类树结构, 交通词汇知识网络具有四点优势: 第一, 打破了抽象词汇和具象词汇的壁垒, 从多种角度发掘词汇的语义联系, 第二, 领域内的词汇意义明确, 避免了一词多义的现象; 第三, 可以方便地根据实际应用的要求动态增减领域词汇的规模; 第四, 能够快速推广到其它领域的词汇.

2.2 基于平均路径长度的相似度计算

基于信息内容的相似度算法需要额外的语料库支撑, 不利于海量数据的计算. 本文基于路径长度计算词汇的语义相似度. 与 wordnet 中单一的上下义关系不同, 领域词汇网络中每种“线形”的路径都代表了一种在 2.1 节中新加入的语义关系, 如果仍然以最短路径计算相似度, 将忽略词汇间多元的语义联系, 不符合人工语义判断的规律. 式(3)中的路径长度不再是节点间的最短路径长度, 而是由节点间各类型的路径长度平均得到. 此外, 基于平均路径长度的算法使得路径长度参数的取值范围由整数扩大为实数, 计算的语义相似度将粒度更细, 更加平滑. 计算公式如下:

$$\text{wordsim}(w_1,w_2)=-\log\left(\frac{\overline{len(w_1,w_2)}}{2*D}\right) \tag{8}$$

$$\overline{\text{len}(w_1,w_2)}=\frac{\sum_{i=1}^{l} len_i(w_1,w_2)}{l} \tag{9}$$

其中, $\overline{len(w_1,w_2)}$ 为 w_1 和 w_2 的平均路径长度, D 为 wordnet 中最深节点层数, $len_i(w_1,w_2)$为第 i 种类型路径的长度. L 位路径总数.

3　实验结果及分析

3.1 实验设置与评价指标

结合图像标签排序的实际应用需求, 从图像分享网站 Flickr 上下载了带有“traffic”或者“vehicle”标签的图像1000幅. 预处理后, 这1000幅图像一共包含2016个不重复的标签, 我们将出现次数排在前 300 的标签作为构造领域知识网络的常用词汇, 具体包括 traffic、vehicle、car、people、street 等. 为了直观评价相似度的计算结果, 分别采用基于路径长度的 Wu 和 Palmer 算法(以下称 WP 算法)、基于信息内容的 Jiang 和 Conrath 算法(以下称 JC 算法)以及本文算法来计算“vehicle”与其它交通词汇的语义相似度. 图 5 展示了归一化后的相似度变化曲线:

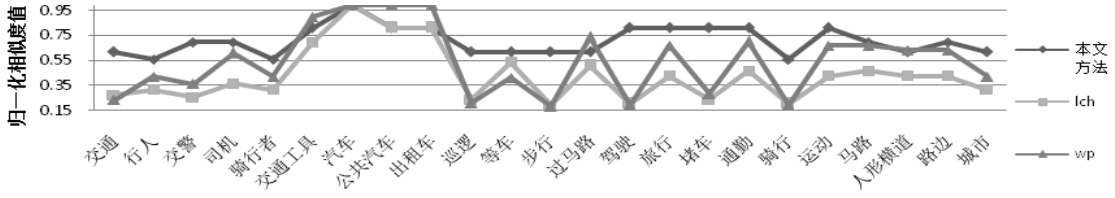


图 5　“vehicle”与部分交通词汇的语义相似度曲线

由图 5 可知, 1)WP 算法和 JC 算法的相似度曲线除了偏置量的差别, 变化趋势基本一致; 2)本文算法由于引入了多种词汇关系, 领域知识的相似度得到加强, 整体高于 WP 和 JC 算法; 3)与交通主体和交通地点中的词汇相比, WP 和 JC 算法中交通行为词汇(多为抽象事物)与“汽车”的相似度值偏小, 而本文算法采用的关系主体-事件和场所-事件等关系克服了抽象事物和具象事物之间的天然“弱相似性”, 使结果明显增大.

Finkelstein 给出了一个包含 353 对词汇的语义相似度测试集, 测试集中的每对词汇都是由专家精心挑选, 涵盖了从“高语义相关”到“语义不相关”的类型. 为了得到真实的人工评价, 邀请了 51 个受试者相互独立的对这 353 对词汇的“意义相似性”进行打分, 分值从 0.0 到 4.0 变化. 受试者打分的平均值即为该测试集的真实值. 皮尔森关联度[13]是评价一个词汇相似度算法的好坏常用标准. 它反映了算法所得的相似度值和 Finkelstein 测试集中人工判断的结果的符合程度, 关联度越高, 算法越好. 计算公式如下:

$$\mathrm{Cor}(x,y)=\frac{\sum_i\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum_i\left(x_i-\bar{x}\right)^2}\sqrt{\sum_i\left(y_i-\bar{y}\right)^2}} \tag{10}$$

从 Finkelstein 测试集中选取了 56 对交通相关的词汇, 采用 wp 算法、resnik 算法、Lch 算法、Lin 算法和本文算法计算相似度, 各算法的皮尔森关联度和部分计算结果如图 6 和表 2 所示.

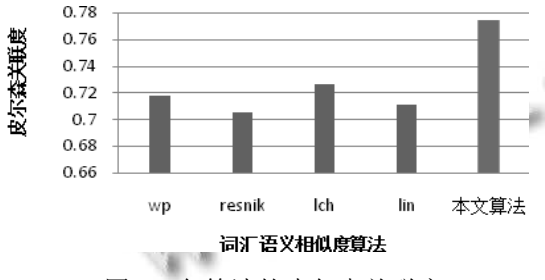


图 6 各算法的皮尔森关联度

表 2 部分 Finkelstein 词汇对不同算法的相似度值

词汇对	真实值	Lch 算法	本文算法
car-driveing	0.79	0.37	0.7
car-journey	0.64	0.19	0.56
cushion-automobile	0.24	0.40	0.17
Automobile-car	0.98	1	1
vehicle- underground	0.91	0.02	0.81
steering-vehicle	0.67	0.32	0.75
track- train	0.82	0.56	0.64
street-highway	0.78	0.62	0.69

整体而言, 本文算法的计算结果更接近真实值, 有着更高的皮尔森关联度值. 如表 2 所示, 在抽象词汇与具体词汇对(比如 car-journey 和 car-driving 等)的相似度计算中, 本文算法的结果明显优于 lch 算法. 而在 lch 算法中关系不大的 steering 和 vehicle, 由于部件-整体关系的引入, 本文算法获得了更符合人工判断的相似度值.

3.2 在图像标签排序中的应用

图像标签排序是根据标签与图像内容的相关程度由大到小重新排列标签. 词汇语义相似度体现了标签之间的亲疏关系, 是图像标签排序的重要依据. 从新加坡国立大学提供的 NUS-WIDE 测试集[14]中选取了“traffic”类别的图像 300 幅用于标签排序. 语义相似度分别由 Lin、Lch 和本文算法得到. 实验采用归一化折损累积增益值(Normalized Discounted Cumulative Gain, NDCG)作为评价指标. 实验前, 由志愿者基于标签与图像的相关度, 对测试集中每个标签进行打分, 分值共有 5 个等级, 为 0 至 4 的整数, 数值越大, 相关度越大. 图像标签的 NDCG 值的计算公式如下:

$$s_{NDCG}=z\sum_{i=1}^{n}(2^{r(i)}-1)/\log(1+i) \tag{11}$$

其中 Z 是在最佳排序时, 使 NDCG 值归一化为 1 的系数, $r(i)$表示第 i 个标签的相关度得分. 图 7 为各种算法取得的平均 NDCG 值. 图 8 展示了排序前后的标注情况.

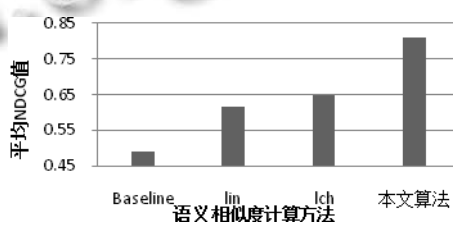
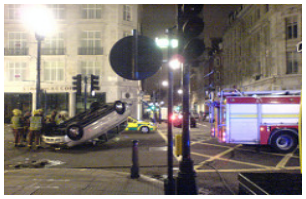


图 7 原始标签和各种算法排序后的 NDCG 值

由图可知, 原始标注的顺序很随意, 一些与图像内容无直接关系的标签往往占据着靠前的位置. 三种方法都不同程度地改善了标签的排列顺序. 相比于 Lin 算法, 本文算法在抽象词汇与具象词汇相似度计算上的修正, 使得诸如“traffic-jam”、“car-crash”和“accident”等表示交通行为的词汇得到“重视”, 排到了前列, 取得了更高的 NDCG 值. 这些词汇连同其它表

示交通主体、交通地点的词汇一起立体地描绘了图像中的“交通场景”.

原始标签： china city cars downtown shanghai traffic-jam car overpass traffic rush-hour
Lin 法： city shanghai downtown china cars car traffic overpass rush-hour traffic-jam
本文方法： cars car traffic rush-hour overpass city shanghai downtown china traffic-jam

原始标签： street traffic-light London oxford car crash accident starbucks reportage pavement taxi whack vehicle firemen car-crash
Lin 法： car vehicle taxi street pavement oxford London car-crash crash accident traffic-light whack starbucks firemen reportage
本文方法： car vehicle taxi car-crash crash accident traffic-light street pavement oxford London firemen whack starbucks reportage

图 8　部分图像的标签排序结果

4　结语

词汇的语义相似度计算是信息检索和图像标签处理等中的基本问题. 常用的计算方法有基于 wordnet 树结构的路径长度法和信息内容法. 针对传统算法的不足, 本文提出了基于交通领域知识网络的词汇语义相似度算法. 在上下义关系的基础上, 增加了部件-整体、属性-宿主和工具-工具对象等 9 种关系将交通领域内的常用词汇重新构造为互相联系的知识网络. 网络中的每条通路都代表了通路上节点的一种语义关联, 基于这些通路的平均路径长度, 我们定义了一种新的词汇的语义相似度算法. 在 Finkelstein 测试集和 NUS-WIDE 图像集上的实验表明, 本文算法可以取得更符合人工判断的词汇语义相似度.

参考文献

1 Pedersen T, Pakhomov SVS, Patwardhan S, et al. Measures of semantic similarity and relatedness in the biomedical domain. Journal of Biomedical Informatics, 2007, 40(3): 288–99.

2 孙叔琦.基于统计的词汇语义相关计算研究[博士学位论文].哈尔滨:哈尔滨工业大学,2014.

3 Mohammad SM, Hirst G. Distributional measures of semantic distance: A survey. Computer Science, 2012.

4 Adhikari A, Singh S, Dutta A. A novel information theoretic approach for finding semantic similarity in WordNet. TENCON. Macao, China. IEEE. 2015. 1–6.

5 Harispe S, Ranwez S, Janaqi S, et al. Semantic measures for the comparison of units of language, concepts or instances from text and knowledge base analysis. Computer Science, 2013.

6 Hoffart J, Seufert S, Nguyen D B, et al. KORE: Keyphrase overlap relatedness for entity disambiguation. 21st ACM International Conference on Information and Knowledge Management. CIKM. NY, USA. ACM. 2012. 545–554.

7 Hirst G, St-Onge D. Lexical chains as representations of context for the detection and correction of malapropisms. Fellbaum C. WordNet: An Electronic Lexical Database. Cambridge, MA, USA: MIT Press, 1998: 305–332.

8 Yu X, Sun Y, Norick B. User guided entity similarity search using meta-path selection in heterogeneous information networks. Proc. of the 21st ACM International Conference on Information and Knowledge Management. NY, USA. ACM. 2012. 2025–2029.

9 王桐,王磊,吴吉义.wordnet 中的综合概念语义相似度计算方法.北京邮电大学学报,2013,36(2): 98–101.

10 Lin D. An information-theoretic definition of similarity. Proc. of the Fifteenth International Conference on Machine Learning. San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. 1998. 296–304.

11 Formica A. Concept similarity in formal concept analysis, Information Science, 2006, 176(18): 2624–2641.

12 董强,董振东.知网简介. http://www.keenage.com/.

13 刘宏哲,须德.基于本体的语义相似度和相关度计算研究综述.计算机科学,2012,39(2): 8–13.

14 Chua TS, Tang J, Hong R, et al. NUS-WIDE: A real-world web image database from National University of Singapore. ACM International Conference on Image and Video Retrieval. ACM. 2009. 1–9.