文章编号: 1671-5896(2010)06-0602-07

# 基于同义词词林的词语相似度计算方法

田久乐, 赵 蔚
(东北师范大学 计算机科学与信息技术学院, 长春 130117)

摘要: 为解决词语相在语义网自适应学习系统中相似度计算不清的问题, 以同义词词林为基础, 提出并实现了一种基于同义词词林的词语相似度计算方法, 充分分析并利用了同义词词林的编码及结构特点。该算法同时考虑了词语的相似性, 和词语的相关性。进行人工测试, 替换测试以及与当前流行的基于“知网”的词语相似度算法对比测试的结果表明, 该算法与人们思维中的相似度值基本一致, 有较高的准确性。

关键词: 词语相似度; 同义词词林; 自适应学习系统
中图分类号: TP391.5 文献标识码: A

# Words Similarity Algorithm Based on Tongyici Cilin in Semantic Web Adaptive Learning System

TIAN Jiu-le, ZHAO Wei
(School of Computer Science and Information Technology, Northeast Normal University, Changchun 130117, China)

**Abstract**: Words similarity has a role which cannot be ignored in the semantic web adaptive learning system. We propose and implement words similarity algorithm based on Tongyici Cilin, in which we fully analyze and use the coding and structural characteristics of Tongyici Cilin. We consider both the words resemblance and the words relevance. After manual test, replacement test and comparison test in which we compared our algorithm with the current popular words similarity algorithm based on Hownet, we found that the algorithm achieved the good results.

**Key words**: word similarity; tongyici cilin; adaptive learning system

## 引 言

网络教育是开放式的教育体系。它以学习者为主体, 打破了传统教育中时空和地域的限制, 学习者可根据自己的学习计划在网上学习, 并充分自由地利用教育资源[1]。目前, 在我国, 网络教学平台在展现课程基本内容、发布教学信息方面已经发展得比较成熟。但目前的网络课程基本上还是把现有的资源按照一定的顺序摆放在网络上, 学习者进入后只能按部就班地对知识进行接受, 这种网络课程不能根据学习者的认知特征和知识背景动态地呈现最适合学习者学习的内容[2]。因此, 解决学生需求的个性化与教学资源的静态化的最有效方案是构建自适应学习系统。在自适应学习的条件下, 学习不是被动接受知识的过程, 而是主动发现知识的过程。学生能自我组织、制订并执行学习计划, 并能控制整个学习过程, 对学习进行自我评估[3]。要想实现这些功能, 语义网无疑是最佳的平台。语义网克服了传统网络无法理解语言逻辑意义的缺点, 基于语义网构建的自适应学习系统, 为学生的学习提供了非常有效的支持工具。要实现语义网的基本功能, 如, 资源推荐、网页的语义标注、语义搜索引擎、自然语言问答

收稿日期: 2010-09-26
基金项目: 教育部人文社会科学规划基金资助项目(08JA880012); 吉林省科技发展计划基金资助项目(20070521)
作者简介: 田久乐(1988— ), 女, 吉林磐石人, 东北师范大学硕士研究生, 主要从事个性化学习系统研究, (Tel) 86-15590555998(E-mail) tianjl261@nenu.edu.cn; 赵蔚(1963— ), 女, 长春人, 东北师范大学计算机科学与信息技术学院副院长, 教授, 博士生导师, 主要从事个性化学习系统研究, (Tel) 86-13353106026(E-mail) zhaow577@nenu.edu.cn.

以及语义冲突的消解等, 词语的相似度计算则成为基础。

## 1 词语的相似度计算

词语相似度计算在自然语言处理、智能检索、文本聚类、文本分类、自动应答、词义排歧和机器翻译等领域都有广泛的应用, 它是自然语言的基础研究课题, 正在被越来越多的研究人员所关注[4]。

目前, 国内外词语相似度计算方法之一是基于语义词典的词语相似度计算。常用的语义词典: 在英文方面, 具有代表性的有 WordNet, FrameNet, MindNet等; 在汉语方面, 有 “知网” (HowNet)、“同义词词林”“中文概念词典” (CCD: Chinese Concept Dictionary) 等。该算法即根据同义词词林的编排及语义特点计算两个词语之间的相似度。利用本算法提出的相似度与人们思维习惯中的词语相似度非常接近, 而且可以方便、全面地找出替换词语。

## 2 同义词词林介绍

《同义词词林》是梅家驹等人[5]于 1983年编纂而成, 这本词典中不仅包括了一个词语的同义词, 也包含了一定数量的同类词, 即广义的相关词。由于 《同义词词林》著作时间较为久远, 且之后没有更新, 所以哈尔滨工业大学信息检索实验室利用众多词语相关资源, 完成了一部具有汉语大词表的 “哈工大信息检索研究室同义词词林扩展版”。《同义词词林扩展版》收录词语近 7万条, 全部按意义进行编排, 是一部同义类词典。哈工大信息检索研究室参照多部电子词典资源, 并按照人民日报语料库中词语的出现频度, 只保留频度不低于 3的 (小规模语料的统计结果) 部分词语, 剔除 14 706个罕用词和非常用词后, 词表共包含 77 343条词语。

### 2.1 同义词词林分类方法

同义词词林按照树状的层次结构把所有收录的词条组织到一起, 把词汇分成大、中、小 3类, 大类有 12个, 中类有 97个, 小类有 1 400个。每个小类里都有很多的词, 这些词又根据词义的远近和相关性分成了若干个词群 (段落)。每个段落中的词语又进一步分成了若干个行, 同一行的词语要么词义相同 (有的词义十分接近), 要么词义有很强的相关性。例如, “大豆”、“毛豆” 和 “黄豆” 在同一行; “西红柿” 和 “番茄” 在同一行; “大家”、“大伙儿”、“大家伙儿” 在同一行。另外, “将官”、“校官”、“尉官” 在同一行, “雇农”、“贫农”、“下中农”、“中农”、“上中农”、“富农” 在同一行, “外商”、“官商”、“坐商”、“私商” 也在同一行, 这些词不同义, 但很相关。

同义词词林词典分类采用层级体系, 具备 5层结构, 如图 1所示。随着级别的递增, 词义刻画越来越细, 到了第 5层, 每个分类里词语数量已经不大, 很多只有一个词语, 已经不可再分, 可以称为原子词群、原子类或原子节点。不同级别的分类结果可以为自然语言处理提供不同的服务, 例如第 4层的分类和第 5层的分类在信息检索、文本分类、自动问答等研究领域得到应用。研究证明, 对词义进行有效扩展, 或对关键词做同义词替换可以明显改善信息检索、文本分类和自动问答系统的性能。

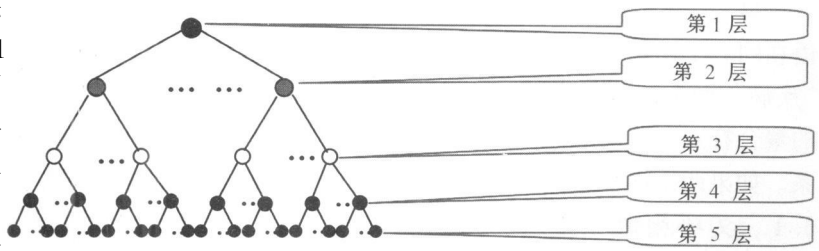


图 1 同义词词林 5层结构

Fig. 1 Five-layer structure of Tongyici Cilin

### 2.2 同义词词林编码方法

同义词词林共提供了 5层编码, 第 1级用大写英文字母表示; 第 2级用小写英文字母表示; 第 3级用二位十进制整数表示; 第 4级用大写英文字母表示; 第 5级用二位十进制整数表示。例如: “Ae07C01= 渔民 渔翁 渔家 渔夫 渔父 打鱼郎”, “Ae07C01= ” 是编码, “渔民 渔翁 渔家 渔夫 渔父 打鱼郎” 是该类的词语。例如:

Ba01A02= 物质 质 素;
Cb02A01= 东南西北 四方;
Ba01A03@ 万物;
Cb06E09@ 民间;
Ba01B08# 固体 液体 气体 流体 半流体;
Ba01B10# 导体 半导体 超导体。
具体编码如表 1所示。

表 1 词语编码表

Tab. 1 The table of code of word

| 编码位 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 符号举例 | D | a | 1 | 5 | B | 0 | 2 | = \# \@ |
| 符号性质 | 大类 | 中类 | 小类 | | 词群 | 原子词群 | | |
| 级别 | 第 1级 | 第 2级 | 第 3级 | | 第 4级 | 第 5级 | | |

由于第 5级有的行是同义词, 有的行是相关词, 有的行只有一个词, 分类结果需要特别说明, 可以分出具体的 3种情况。在使用上, 有时需要对这 3种情况进行区别对待, 所以有必要再增加标记分别代表几种情形。表 1中的编码位是按照从左到右的顺序排列。第 8位的标记有 3种, 分别是 “=”、“#”、“@”。“=”代表 “相等”、“同义”; “#”代表 “不等”、“同类”, 属于相关词语;。“@”代表 “自我封闭”、“独立”, 它在词典中既没有同义词, 也没有相关词。

## 3 基于同义词词林的词语相似度算法

词语相似度是个数值, 一般取值范围在 [0, 1] 之间。一个词语与其本身的语义相似度为 1。如果两个词语在任何上下文中都不可替换, 则其相似度为 0。相似度涉及到词语的词法、句法、语义甚至语用等方面的特点。影响词语相似度的两个重要指标是词语相似性和词语相关性。其中影响最大的是词语的相似性。词语相似性由人为判断, 所以具有较强的主观性。一般用词语间的语义距离 (词语距离) 衡量词语的相似度。具体地说, 词语距离是一个 [0, ∞) 之间的实数。一个词语与其本身的距离为 0。词语距离与词语相似度之间有密切关系。两个词语的距离越大, 其相似性越低; 反之, 两个词语的距离越小, 其相似性越高[6]。

另一个影响词语相似度的是词语的相关性, 它反映的是两个词语互相关联的程度。可以用这两个词语在同一个语境中共现的可能性衡量。词语相关性也是一个 [0, 1] 之间的实数。例如 “医生”和 “疾病”两个词语, 其相似性非常低, 而相关性却很高。可以这么认为, 词语相似性反映的是词语之间的聚合特点, 而词语相关性反映的是词语之间的组合特点。同时, 词语相关性和词语相似性又有密切联系。如果两个词语相似很高, 则这两个词语也会有很高的相关性[7]。

### 3.1 义项相似度基础算法

中文词语博大精深, 一个词语往往表达了很多的意思, 也就是说有很多个义项。例如 “骄傲”, 既可以表示褒义: 自豪, 也可以表示贬义: 傲慢。因此计算词语相似度要考虑到所有的义项。

该算法根据同义词词林编排特点, 提出基于同义词词林的义项相似度的主要思想是: 基于同义词词林结构, 利用词语中义项的编号, 根据两个义项的语义距离, 计算出义项相似度。算法如下。

首先判断在同义词林中作为叶子节点的两个义项在哪一层分支, 即两个义项的编号在哪一层不同。例如: Aa01A01与 Aa01B01, 即在第 4层分支。从第 1层开始判断, 相同则乘 1, 否则在分支层乘以相应的系数, 然后乘以调节参数 $\cos\left(n\times\frac{\pi}{180}\right)$, 其中 $n$ 是分支层的节点总数, 该调节参数的功能是把义项相似度控制在 [0, 1] 之间。

词语所在树的密度, 分支的多少直接影响到义项的相似度, 密度较大的义项相似度的值相比密度小的相似度的值精确。再乘以一个控制参数 $(n-k+1)/n$, 其中 $n$ 是分支层的节点总数, $k$ 是两个分支间

的距离。这样把原本计算出的只对应在几点的值细化, 精确计算结果。

若两个义项的相似度用 Sim表示

1) 若两个义项不在同一棵树上

$$\mathrm{Sim}(A, B) = f; \tag{1}$$

2) 若两个义项在同一棵树上:

①若在第 2层分支, 系数为 $a$

$$\mathrm{Sim}(A, B) = 1 \times a \times \cos\left(n \times \frac{\pi}{180}\right)\left(\frac{n - k + 1}{n}\right) \tag{2}$$

②若在第 3层分支, 系数为 $b$

$$\mathrm{Sim}(A, B) = 1 \times 1 \times b \times \cos\left(n \times \frac{\pi}{180}\right)\left(\frac{n - k + 1}{n}\right) \tag{3}$$

③在第 4层分支, 系数为 $c$

$$\mathrm{Sim}(A, B) = 1 \times 1 \times 1 \times c \times \cos\left(n \times \frac{\pi}{180}\right)\left(\frac{n - k + 1}{n}\right) \tag{4}$$

④在第 5层分支, 系数为 $d$

$$\mathrm{Sim}(A, B) = 1 \times 1 \times 1 \times 1 \times d \times \cos\left(n \times \frac{\pi}{180}\right)\left(\frac{n - k + 1}{n}\right) \tag{5}$$

举例说明如下:

Ae05A01= 邮递员 邮差 信差 信使 绿衣使者 通信员 投递员

Ae05A02= 交通员 交通 通讯员

Ae05A03= 联络员 联络官 联系人

义项邮递员和联络员的相似度为

$$\mathrm{Sim}(A, B) = d \times \cos\left(n \times \frac{\pi}{180}\right)\left(\frac{n - k + 1}{n}\right) \tag{6}$$

其中 $n = 3$, $k = 2$ (邮递员在 01分支, 联络员在 03分支)。

该算法只考虑了分支处的处理, 然后计算相似度的值, 分层之后的层次编号不予考虑。

如计算的两个义项的编号相同, 即在同一行内, 则考虑用编号计算义项的相似度: 当编号相同且末尾号为 = 时, 相似度为 1; 当编号相同而只有末尾号为#时, 直接把定义的系数 $e$ 赋给结果。例如:

Ad02B04# 非洲人 亚洲人

则义项 “非洲人” 和 “亚洲人” 的相似度

$$\mathrm{Sim}(A, B) = e \tag{7}$$

当编号的尾号为@时, 则代表这个词既没有同义词也没有相关词, 在一个编号中只有一个词, 所以不予考虑。

在经过多次试验, 人工评定后将层数初值设置为 $a = 0.65$, $b = 0.8$, $c = 0.9$, $d = 0.96$, $e = 0.5$, $f = 0.1$。

### 3.2 词语相似度算法

在计算词语相似度时, 把两个词语的义项分别两两计算, 取最大值作为两个词语的相似度值。

例如: 词语 “骄傲” 的义项编号有 “Da13A01= ”、 “Ee34D01= ”; 词语 “仔细” 的义项编号有 “Ee26A01= ”、 “Ee28A01= ”。

分别计算义项的相似度为: 0.1, 0.1, 0.483 920, 0.510 077。可以得出 “骄傲” 和 “仔细” 的词语相似度为 0.510 077, 即 4个义项相似度的最大值。

# 4 测 试

## 4.1 人工测试

对实验结果进行人工评价，评价方法是对计算机得到的词语相似度和人对词语相似的排列结果进行比较。结果表明，笔者所用算法的计算结果和人工按照语义相似度的排序结果基本一致。

对同义词词林中100个词语做相似度计算，以其中一个词语“人民”为例，结果如表2所示。

表2 词语“人民”与其他词语的语义相似度

Tab.2 The similarity between word “renmin” and other words

| 词语 | 词语相似度 | 相似度范围 |
|---|---|---|
| 国民 | 1.000 000 | 1.0 |
| 群众 | 0.957 661 | 0.9~1.0 |
| 党群 | 0.897 808 | 0.8~0.9 |
| 良民 | 0.718 246 | 0.7~0.8 |
| 同志 | 0.663 015 | 0.6~0.7 |
| 成年人 | 0.582 177 | 0.5~0.6 |
| 市民 | 0.499 009 | 0.4~0.5 |
| 亲属 | 0.360 396 | 0.3~0.4 |
| 志愿者 | 0.207 921 | 0.2~0.3 |
| 先锋 | 0.166 336 | 0.1~0.2 |

可以看出，语义相似度的计算结果与实际人工判断的语义相似度基本一致，能真实地反应客观现实。实验结果表明，该算法能准确客观地反映词语之间的语义关系，为词语间的语义关系提供一种有效度量。

## 4.2 非人工测试

在非人工测试中，主要进行了以下两方面的测试。

### 4.2.1 替换测试

替换测试是根据计算出的相似度把词语从句子中替换出来，人工评定替换后的句子能否表达原来的意思。

按照一定规则抽取50个例句，提取出主、谓、宾关键词，给定相似度值（1/0.9/0.8），用同义词林找出符合该相似度的词，人工判断这些词在原句中的可替换性。例如对句子“我喜欢学习”的替换(见表3)。

表3 基于同义词词林的相似度计算替换表

Tab.3 Replacement table based on Tongyici Cilin

| | 相似度 | 词语编号 | 可替换词语 |
|---|---|---|---|
| 主语替换 | 1.000 000 | Aa02A01 | 我 咱 俺 余 吾 予 侬 咱家 本人 个人 人家 |
| | 0.935 395 | Aa02A02 | 鄙 愚 鄙人 小人 小子 在下 |
| | 0.899 452 | Aa02B01 | 咱 |
| 谓语替换 | 1.000 000 | Gb09A01 | 喜欢 喜爱 喜好 爱慕 爱好 希罕 好 爱 喜 嗜 |
| | 0.899 452 | Gb09B01 | 爱 喜爱 钟爱 热爱 酷爱 挚爱 热衷 |
| | 0.899 452 | Gb09B02 | 恋 爱恋 热恋 |
| | 0.899 452 | Gb09B0 | 偏好 |
| 宾语替换 | 1.000 000 | Hg08A01 | 学习 念书 读书 上学 就学 求学 修业 攻读 深造 读念 学 习 攻修 |
| | 0.898 767 | Hg08B01 | 从师 投师 拜师 受业 执业 |
| | 0.883 685 | Hg08A02 | 复习 温习 温课 复课 习 温书 |
| | 0.806 843 | Hg08A04 | 自学 自习 自修 进修 |

表3中显示的可以替换的词语。可见，该算法可替换的词语数量多而且精确，都在人主观可以接受的范围内。

### 4.2.2 对比测试

目前，语义词典的计算大多数是基于知网的，所以将该算法与基于知网的词语相似度算法进行对比测试[8]（见表4，表5）。分别在同义词林和知网的词典中，任意抽取100个词，计算并打印出与每个词的相似度大于或等于1/0.9/0.8的所有词。人工判断该级别词与相邻级别词的准确度。

表 4　基于同义词词林的部分计算结果

Tab. 4　Calculation result based on Tongyici Cilin

| 词语相似度 | 编　号 | 词　语 | 相似度范围 |
|---|---|---|---|
| 1.000 000 | Dc02A01= | 风景 景色 景致 景 色 山水 风光 光景 风月 风物 景物 景点 景观 山色 青山绿水 山山水水 山光水色 | 1.0 |
| 0.935 395 | Dc02A02= | 春光 春色 韶光 韶华 蜃景 春暖花开 | 0.9~ 1.0 |
| 0.863 442 | Dc02A03= | 秋景 秋色 | 0.8~ 0.9 |
| 0.899 452 | Dc02B01= | 景象 现象 气象 状况 情景 光景 面貌 场面 容 观 景 万象 场景 此情此景 形貌 | |
| | Dc02B02= | 美景 良辰美景 | |
| | Dc02B03= | 幻境 幻影 春梦 幻梦 镜花水月 幻景 幻像 | |
| | Dc02B06= | 村容村貌 村容 | |
| | Dc02B07= | 新气象 新景观 新貌 | |
| | Dc02B08# | 奇观 壮观 旧观外观 奇景 别有天地 | |
| | Dc02B09# | 镇容 矿容 市容 院容 | |
| | Dc02B10# | 市貌 院貌 | |

表 5　基于知网的计算结果

Tab. 5　Calculation result based on Hownet

| 词语相似度 | 词　语 | 相似度范围 |
|---|---|---|
| 1.000 000 | 色 山水 山光水色 山色 山山水水 风月 视线 风貌 风 风光 风景 观 观瞻 光景 湖光山色 景色 镜头景 景点 景观 景物 景致 图景 | 1.0 |
| 0.950 000 | 前景 地 底 白底 | 0.9~ 1.0 |
| 0.948 000 | 夜景 春光 | |
| 0.923 000 | 背景 | |
| 0.905 609 | 街景 | |
| 0.879 630 | 年景 | 0.8~ 0.9 |
| 0.807 692 | 面貌 面目 | |
| 0.800 000 | 暮色 美景 奇景 奇观 全景 大观 风物 | |

表 4, 表 5显示了分别用同义词词林和知网对同一个词语相似度计算的对比结果。可见两个软件的计算结果基本一致, 但两表中处于 0.9~ 1之间的次差别较大, 但对计算结果影响不大。

### 4.2.3　结果分析

以上的测试办法从不同角度验证了基于同义词词林的词语相似度计算方法的准确性和合理性。总结该算法优点是:

1) 词语相似度的计算结果与人们思维中的相似度值基本一致, 说明该算法的计算结果具有较高的准确性;

2) 对句子的词语替换测试表明, 该算法不但可以较全面地找出可替换词语, 并且替换后的词语仍然可以表达句子原有的意思;

3) 在与基于知网的词语相似度计算结果的对比表明, 不同的语义词典的结果组织不同, 结果的侧重也不同, 所以结果有一定的差距, 但相似度计算结果基本一致。

## 5　结　语

笔者提出了一种基于同义词词林的词语相似度的计算方法, 该算法从词语的语义出发, 根据词语的义项在同义词词林的位置和编码, 计算出词语的相似度。测试证明, 该算法与人们思维中的相似度值基本一致, 可以准确而全面地找出替换词语。

在自适应学习系统中, 将利用该算法进一步研究短语之间的相似度以及句子之间的相似度, 用来实现和完善基于语义网的自适应学习系统的基本功能。

参考文献:

[1]赵蔚, 刘秀琴, 邱百爽. 语义网自适应学习系统中领域本体的构建 [J]. 吉林大学学报: 信息科学版, 2008, 26(5): 514-518.
ZHAO Wei, LIU Xiu-qin, QIU Bai-shuang. Construction of Domain Ontology in Adaptive Learning System Based on Semantic Web [J]. Journal of Jilin University: Information Science Edition, 2008, 26(5): 514-518.

[2]YU Sheng-quan, HE Ke-kang. The Research of Adaptive Learning System Based on Internet [C] // The Third Global Chinese Computer Application Conference Analects. Macao, China: Macao University Press, 1999: 34-40.

[3]邱百爽, 赵蔚, 刘秀琴. 基于语义网的自适应学习系统中用户模型的研究 [J]. 开放教育研究, 2008, 14(4): 106-111.
QIU Bai-shuang, ZHAO Wei, LIU Xiu-qin. The Research of User Model in Adaptive Learning System Based on Semantic Web [J]. Open Education Research, 2008, 14(4): 106-111.

[4]余刚, 裴仰军, 朱征宇, 等. 基于词汇语义相似度计算的文本相似度研究 [J]. 计算机工程与设计, 2006, 27(2): 241-244.
YU Gang, PEI Yang-jun, ZHU Zheng-yu, et al. Research of Text Similarity Based on Word Similarity Computing [J]. Computer Engineering Design, 2006, 27(2): 241-244.

[5]梅家驹, 竺一鸣, 高蕴琦, 等. 同义词词林 [M]. 上海: 上海辞书出版社, 1993: 106-108.
MEI Jia-ju, ZHU Yi-ming, GAO Yun-qi, et al. Tongyici Cilin [M]. Shanghai: Shanghai Lexicographical Publishing House, 1993: 106-108.

[6]关毅, 王晓龙. 基于统计的汉语词汇间语义相似度计算 [C] // 语言计算与基于内容的文本处理——全国第七届计算语言学联合学术会议论文集. 北京: 清华大学出版社, 2003: 221-227.
GUAN Yi, WANG Xiao-long. Similarity between Chinese Words Based on Statistics [C] // Language Computing and Text Processing Based on Content—JSCL2003. Beijing: Tsinghua University Press, 2003: 221-227.

[7]程涛, 施水才, 王霞, 等. 基于同义词林的中文文本主题词提取 [J]. 广西师范大学学报: 自然科学版, 2007, 25(2): 145-148.
CHENG Tao, SHI Shui-cai, WANG Xia, et al. Thematic Words Extracting from Chinese Text Based on Tongyici Cilin [J]. Journal of Guangxi Normal University: Natural Science Edition, 2007, 25(2): 145-148.

[8]刘群, 李素建. 基于“知网”的词汇语义相似度计算 [C] // 计算语言学与中文语言处理——第三届汉语词汇语义学研讨会论文集. 台北: 台北市中研院语言学研究所, 2002: 59-76.
LIU Qun, LI Su-jian. Word Similarity Computing Based on How-Net [C] // Computational Linguistics and Chinese Language Processing—Proceedings of the Third Chinese Lexical Semantics Workshop. Taipei: Institute of Linguistics, Academia Sinica, 2002: 59-76.

(责任编辑: 刘东亮)