

知网的形成概念分析及概念相似度研究^{*}

廖 剑, 冷 静, 李艳燕[†], 黄荣怀
(北京师范大学 知识科学与工程研究所, 北京 100875)

摘 要: 将知网的义项、义原及其关系映射到形式化概念分析的语境中, 生成一个基于知网的形成概念格。一方面, 提供了一种将知网中概念关系转换为概念格的表征方式, 从格中任意一个节点出发, 可以很方便地访问到与此相关的各种知识, 从而为信息检索和知识推理提供很大方便; 另一方面, 也提出了一种通过对形成概念格进行分析来计算概念相似度的方法。实验证明该方法克服了以往计算方法的若干不足, 并能有效地在相关应用领域如协作学习言论分析中加以应用。

关键词: 知网; 形式概念分析; 概念相似度

中图分类号: TP39 文献标志码: A 文章编号: 1001-3695(2007)11-0032-05

Research on HowNet based on formal concept analysis and concept similarity

LIAO Jian, LENG Jing, LI Yan-yan[†], HUANG Rong-huai
(Knowledge Science & Engineering Institute, Beijing Normal University, Beijing 100875, China)

Abstract: A formal concept lattice was implemented which based on HowNet through mapping the concept, sememe and the relationship in HowNet into formal context. On the one hand, the formal concept lattice as a representation transformed all the conceptual relationship in HowNet. In other words, from any node of the lattice, various kinds of related knowledge were accessible. Thus it provided much convenience to information retrieval and knowledge reasoning. On the other hand, a method of calculating the concept similarity was proposed by analyzing the formal concept lattice, which had overcome the some shortcomings of former algorithms and been proved to be effective when applied to analyze speech acts of collaborative learning in experimental result.

Key words: HowNet; formal concept analysis; concept similarity

语义分析是自然语言理解的重点和难点。近年来, 一些大规模、可计算的语义知识库, 包括 WordNet^[1]、MindNet^[2, 3]、FrameNet^[4]等的开发和利用, 为进行大规模真实文本的语义分析和理解提供了有利的支持。1999 年初, 由中国中文信息学会常务理事董振东先生所主持开发的知网(HowNet), 为自然语言理解提供了一个新的研究资源^[5]。

知网的根本思想在于将各个词语的词义以义项表示, 再通过标准化了的 1 600 多个义原来描述各个义项; 词语与词语之间的关联就主要通过义原的层级及其他关联关系体现。然而, 知网与其他本体相比存在一个问题, 即没有直接定义概念和概念之间的关系。这样, 包括一些本体描述中最常见的上下位将难以获得。本文在仔细、深入分析知网的基础上, 将知网中隐含的各种语义关系全部提取出来, 并按照其义原层级关系进行扩展, 形成一个在形式概念分析中称之为形式化语境的二维表, 并进一步转换为 Hasse 图; 以此为中心建立出知网中概念与概念、概念与特征以及特征与特征之间的内在联系, 形成一种概念格的信息表示结构。在这个网状表示上, 通过任意一个信息入口, 如义项、义原、关系等, 均可以很方便地访问到与此相关的各种知识, 从而为基于知网的信息检索和知识推理提供很大的方便。

1 知网简介

知网是一个以英汉双语所代表的概念以及概念的特征为基础, 以揭示概念与概念之间以及概念所具有的特性之间的关系为基本内容的常识知识库。知网系统的哲学是: “世界上一切事物(物质的和精神的) 都在特定的时间和空间内不停地运动和变化。它们通常是从一种状态变化到另一种状态, 并通常由其属性值的改变来体现。”因此, 知网的运算和描述的基本单位是万物^[6]。

知网着力描述了概念之间和概念属性之间的各种关系, 主要包括上下位关系、同义关系、反义关系、对义关系、属性—宿主关系、部件—整体关系、材料—成品关系、事件—角色关系。在概念词典中, 概念与概念和特征的关系主要体现在每个记录的概念定义项(DEF 项) 中。在各个特征文件中, 这些关系则体现在特征的层次组织树、必要角色框架和共性特征描述项中。所有这些, 均通过知网提供的知识词典描述语言(KDML) 来实现。有关的详细内容可参阅董振东先生的有关论文^[6, 7]。

2 形式概念分析

形式概念分析是应用数学的一个分支, 它建立在概念和概

收稿日期: 2006-09-30; 修返日期: 2007-09-12 基金项目: 国家自然科学基金资助项目(60402016, 60705023)

作者简介: 廖剑(1978-), 男, 硕士研究生, 主要研究方向为 CSCL、文本挖掘等; 冷静(1982-), 女, 硕士研究生, 主要研究方向为 CSCL 等; 李艳燕(通讯作者) (1975-), 女, 博士后, 主要研究方向为知识网络、e-learning、语义搜索、CSCL 等(liyy1114@ 163. com); 黄荣怀(1965-), 男, 教授, 博导, 主要研究方向为 CSCL、知识科学、知识工程等。

念层次的数字化基础之上。运用形式概念分析的方法, 可以发现、构造和展示由属性和对象构成的概念及其之间的关系。目前, 形式概念分析的方法已经大量运用在概念聚类、数据分析、信息检索、知识发现、本体工程的应用之中^[8]。

2.1 语境和概念

定义 1 一个形式化的语境(context) $k = (G, M, I)$, 包含两个集合(G 和 M) 和一个二元关系(G 与 M 之间的关系 I) 。在语境中, G 中的元素称为对象; M 中的元素称为属性。用 gIm , 或者(g, m) I 来表达对象 g 与属性 m 的关系, 读做“对象 g 具有属性 m ”。

根据定义 1, 可以用矩阵来表示语境。列标题上是对象名, 行标题上是属性名。行 g 与列 m 的交叉表示对象 g 具有属性 m (表 1)。

表 1 语境的矩阵表示

对象	属性				
	m_1	m_2	m_3	...	m_n
g_1	X				X
...		X			
g_n			X		

定义 2 对一个对象集 A , 定义 $A' = \{ m \in M | gIm, \text{对所有的 } g \in A \}$ (即 A 中所有对象共有的属性集合)。

相应地, 对一个属性集 B , 定义 $B' = \{ g \in G | gIm, \text{对所有的 } m \in B \}$ (即包含所有 B 中属性的对象集合)。

定义 3 语境(G, M, I) 的形式概念(formal concept) 是一个集合对(A, B)。其中: $A \subseteq G, B \subseteq M$ 并且 $A' = B, B' = A$ 。 A, B 分别称做概念(A, B) 的外延和内涵。 $\mathfrak{K}(G, M, I)$ 表示语境(G, M, I) 中的所有概念集。

命题 1 如果(G, M, I) 是一个语境, 对象集 $A, A_1, A_2 \subseteq G$, 属性集 $B, B_1, B_2 \subseteq M$, 那么有如下结论:

- (1) $A_1 \subseteq A_2 \Rightarrow A_2' \subseteq A_1'$
- (1) $B_1 \subseteq B_2 \Rightarrow B_2' \subseteq B_1'$
- (2) $A \subseteq A''$
- (2) $B \subseteq B''$
- (3) $A' = A$
- (3) $B' = B$
- (4) $A \subseteq B \Leftrightarrow B \subseteq A' \Leftrightarrow A \times B \subseteq I$

2.2 概念格

定义 4 如果(A_1, B_1) , (A_2, B_2) 均是语境中的概念, 并且 $A_1 \subseteq A_2$, 那么(A_1, B_1) 被称做(A_2, B_2) 的子概念, (A_2, B_2) 则是(A_1, B_1) 的超概念, 记为(A_1, B_1) $\leq$ (A_2, B_2) 。“ $\leq$ ”反映了概念间的层次关系。由层次关系搭构的所有(G, M, I) 的概念记做 $\mathfrak{K}(G, M, I)$, 被叫做概念格(concept lattice)。

比较著名的概念格构造算法有 Ganter ’s next-concept algorithm^[12] 以及 Christian Lindig^[13, 14] 的 fast conceptAnalysis 等。这些算法计算出了格的所有概念以及层次关系, 并能据此绘制出相应的 Hasse 图。

3 知网的形式概念分析

对知网进行形式概念分析可以有效地解决知网中概念关系的直接表征问题, 是一种有效的可视化方法。将知网进行形

式概念分析的关键, 就在于将知网中的概念(义项)、特征(义原) 以及它们之间的关联关系映射到形式概念分析的语境中去。本文所采用的策略如下:

- a) 将知网中的义项看做形式概念分析中的对象;
- b) 将知网中的义原及其义原看做形式概念分析中的属性;
- c) 通过获得知网中每个义项的所有义原(包括第一独立义原、其他独立义原、关系义原和符号义原) , 以及这些义原所有的上位义原来构成形式概念分析中的语境;
- d) 在没有进行词义消歧的操作下, 将每个词语所有的义项都进行形式概念分析。

知网的词义消歧步骤可以参见文献[13] , 这里不再详述。

之所以将每个义原的所有上位义原也加入到语境中, 是因为在知网的义原层次关系中蕴涵了这样一条规则, 即如果一个义项含有一个义原对其进行解释, 那么该义原所有的上位义原也能对该义项进行解释。例如“男人 ”这个词语存在一个义项“N human |人, family |家, male |男 ”。其中义原“human |人 ”是一个对“男人 ”的解释, 则其蕴涵了在“human |人 ”之上的所有义原“AnimalHuman |动物 ”“animal |生物 ”“physical |物质 ”“thing |万物 ”“entity |实体 ”也都是对“男人 ”的解释。

在上述策略下, 本文构造了一个对知网进行形式概念分析的分析器。其流程如图 1 所示。

在生成形式化语境后, 即可据此绘制 Hasse 图。Hasse 图^[12] 是概念格的图形化表示。Hasse 图包含语境中对象、属性之间的关系, 是语境的一个等价表示形式。在特定的语境中, 包含类的继承。通过查看 Hasse 图, 能够容易地发现属性之间的依赖和关联。

以词语“飞机、战斗机、武器 ”为例, 本文在分析器中进行了获取义项、获取义原, 并扩展了义原之后, 生成如下形式化语境, 如表 2 所示。

表 2 由知网生成的形式概念语境

属性 \ 对象	飞机	战斗机	武器	属性 \ 对象	飞机	战斗机	武器
aircraft 飞行器	X	X		thing 万物	X	X	X
vehicle 交通工具	X	X		entity 实体	X	X	X
implement 器具	X	X	X	weapon 武器		X	X
artifact 人工物	X	X	X	military 军		X	
inanimate 无生物	X	X	X	secondaryFeature 二特		X	X
physical 物质	X	X	X	generic 统称			X

根据表 2 所示语境, 可以绘制出如图 2 所示的 Hasse 图。

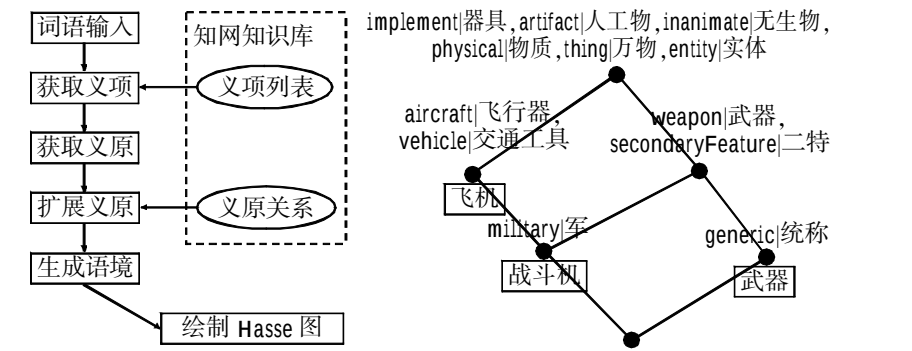


图 1 知网形式概念分析流程

图 2 由知网所生成的形式概念格的 Hasse 图

图 2 中即是一个格 $\mathfrak{K}(G, M, I)$ 。其中节点上显示的文本是形式概念格的属性, 对应知网中的义原。节点下方框中显示

的文本是形式概念格中的对象, 对应知网中的义项。图中圆点则代表形式概念格的概念(注意概念格中概念的含义与知网中的概念含义不同。知网的概念在这里是用义项来代表的)。根据此图可以看到如下性质。

性质 1 对任何一个义项而言, 沿其路径向上, 途中包含自身在内所有节点中的属性集, 为该义项所包含的所有义原集。

性质 2 对任何一个义原而言, 沿其路径向下, 途中包含自身在内所有节点中的对象集, 为该义原所描述的所有义项集。

性质 3 如果一个义项真包含另一个义项所有的义原, 则该义项所在的节点在形式概念格中为另一个义项所在的节点路径之上, 且前一节点为后一节点的子概念, 后一节点为前一节点的超概念。

如图 2 中的“飞机”和“战斗机”。可以看到, “战斗机”包含了“飞机”中所有的义原, 且多出了“military |军”与“weapon |武器”的义原属性。故知网中概念的层级关系可以在形式概念格中很清晰地表达出来。

性质 4 如果一个义项所有的义原与另一个义项所包含的义原相同, 则两个义项在概念格中的同一个节点出现。

通过形式概念分析, 可将知网表征为一种概念格的信息表示结构。在这个网状表示上, 从任意一个信息入口出发, 如义项、义原、关系等, 均可以很方便地访问到与此相关的各种知识, 从而为基于知网的检索与知识推理提供很大方便。此外, 还能据此产生一种新的概念相似度计算方法。

4 知网的相似度计算

4.1 现有知网相似度计算方法

通常对于词语相似度的计算最后均可归为对概念相似度的计算。在文献[16] 所采用的方法中, 对于两个汉语词语 W_1 和 W_2 , 如果 W_1 有 n 个义项(概念) $S_{11}, S_{12}, \dots, S_{1n}$, W_2 有 m 个义项(概念) $S_{21}, S_{22}, \dots, S_{2m}$ 。本文规定, 词语 W_1 和 W_2 的相似度是各个概念的相似度的最大值, 也就是说, $\text{Sim}(W_1, W_2) = \max_{i=1, \dots, n, j=1, \dots, m} \text{Sim}(S_{1i}, S_{2j})$ 。这样, 就把两个词语之间的相似度问题归结到了两个概念之间的相似度问题。

对知网中概念(义项) 的相似度计算通过义原的语义距离来计算^[16~18]。两个义原在这个层次体系中的路径距离为 d , 就可以得到这两个义原之间的语义距离。其公式为 $\text{Sim}(p_1, p_2) = a/(a + d)$ 。其中: p_1 和 p_2 表示两个义原; d 是 p_1 和 p_2 在义原层次体系中的路径长度, 是一个正整数; a 是一个可调节的参数。

对于概念各义原进行比较后的相似度合成, 以实词为例, 其基本计算公式为 $\text{Sim}(S_1, S_2) = \sum_{i=1}^4 \beta_i \text{Sim}_i(S_1, S_2)$ 。其中: $\beta_i (1 \leq i \leq 4)$ 是可调节的参数, 且有 $\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1, \beta_1 \geq \beta_2 \geq \beta_3 \geq \beta_4$ 。 $\text{Sim}_1(S_1, S_2)$ 到 $\text{Sim}_4(S_1, S_2)$ 分别为第一独立义原描述式、其他独立义原描述式、关系义原描述式及符号义原描述式的相似度。

该算法能在一定精度下计算知网中概念的相似度, 然而它

却存在以下问题:

a) 将两个概念的第一独立义原作为权值最大的项进行计算。当两个概念的相同义原存在错位现象时, 则相似度计算上将存在很大误差。以文献[16] 为例, 如比较“飞机”与“战斗机”时, 由上述方法计算语义相似度为 0. 209。然而这两个词语的语义相似程度非常高。这是因为“飞机”的义原为“aircraft |飞行器”, 而“战斗机”的义原为“weapon 武器 |aircraft 飞行器 |military 军”。相同义原存在错位, 由此造成计算上的误差。

b) 没有考虑概念在整个义原层次树中深度的不同对计算结果所产生的影响。以文献[16] 为例, 如比较“友谊”与“特色”时, 语义相似度为 0. 747。从语义上讲, 这两者明显距离较远, 结果却非常相似。分析原因, “友谊”的义原为“attribute |属性, relatedness |亲疏, &human |人”, 而“特色”为“attribute |属性, property |特性, entity |实体”。可以看到, 除第一义原相同外, 余下两个义原均不同, 在只考虑义原上下位关系的情况下, “亲疏”与“特性”路径长度为 3, “人”与“实体”路径长度为 5, 则第二义原相似度为 0. 348, 第三义原相似度为 0. 242。结果却因第一义原相同, 而使最后结果达到了 0. 747 之高, 而第一义原属性处于义原层次结构的最上端, 最后即便两个词语都具备该义原, 并不能说明两词语的相似程度就大(注: 上述结果均在系数 $\alpha = 1. 6, \beta_1 = 0. 5, \beta_2 = 0. 20, \beta_3 = 1. 7, \beta_4 = 1. 3, \gamma = 0. 2, \delta = 0. 2$ 下进行计算)。

由此, 本文基于对知网的形式概念分析提出了一种新的计算方法。

4.2 基于形式概念格的概念相似度计算

在形式概念格中计算概念相似度主要基于以下思想: 在分别描述两个概念的义原集合中, 相同的义原越多, 则证明两个概念越相似。由此, 可以将概念的相似度定义为 $\text{Sim}(S_1, S_2) = (|P_1 \cap P_2| + \alpha|P'_1 \cap P'_2|)/(|P_1 \cup P_2| + \alpha|P'_1 \cup P'_2|)$ 。其中: P_1 为概念 1(义项 1) 所拥有的义原及这些义原在义原层次树中所有祖先义原的集合; P_2 为概念 2(义项 2) 所拥有的义原及这些义原在义原层次树中所有祖先义原的集合; $P_1 \cap P_2$ 为两个概念所共有义原的集合; $P_1 \cup P_2$ 为两个概念所包含的所有义原的集合; P'_1 为概念 1(义项 1) 的第一独立义原及该义原在义原层次树中所有祖先义原的集合; P'_2 为概念 2(义项 2) 的第一独立义原及该义原在义原层次树中所有的祖先义原的集合; $P'_1 \cap P'_2$ 为两个概念所共有义原的集合; $P'_1 \cup P'_2$ 为两个概念所包含的所有义原的集合。引入 P'_1, P'_2 是考虑到知网中第一独立义原对整个概念影响的重要性。其中 α 是第一义原相对于其他义原重要程度的调节因子。其取值为 0 时, 第一独立义原与其他义原有相同重要性; 取值为 1 时, 相当于说明第一独立义原比其他义原更重要一倍。

当 $\alpha = 1$ 时, 等同于在形式概念分析语境中, 将每个概念的第一义原及其层次树上所有的义原再次当做对象的属性, 重复加入到作为语境的二维表格中。由此, 可以根据修正后的语境再次生成概念格的 Hasse 图, 从而对概念的相似度计算提供一种更好的可视化支持。对图 2 进行修正后的概念格如图 3 所示。其中方框中带“^”号标注的属性, 即为每个概念的第一

独立义原及其层次数之上的所有义原。

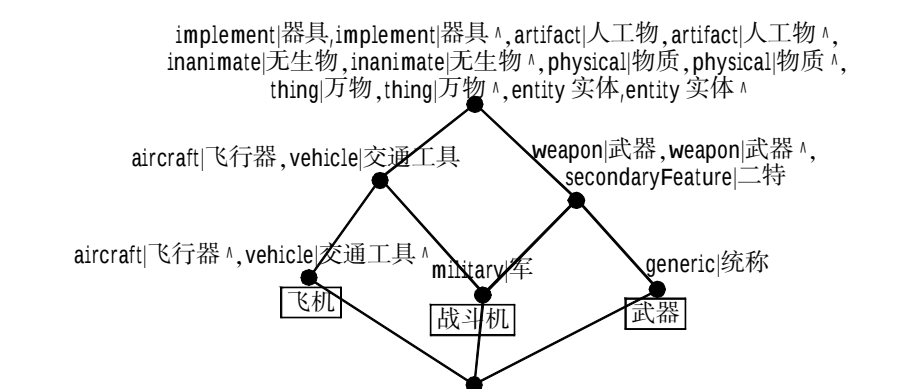


图 3 知网概念相似度计算修正后的形式概念格

通过分析可以发现, 该算法已经很好地避免了文献[17] 中算法所存在的两个问题: a) 通过直接比较相同义原的个数, 避免了在计算两个概念相似度时, 因义项定义中义原位置不对位而造成的计算结果上的误差。b) 义项中所定义的义原在义原层次树中越是靠上, 则由该义原所扩展出来的义原, 即该义原所包含的所有祖先义原也越少, 对最后的计算结果也越小。因此可以有效地避免通过计算路径长度来计算概念相似度的方法中, 义原所在层次树中的深度对计算结果的影响。

4.3 实验研究

以文献[16] 作为对比, 进行以下实验。在尚未对词义进行消歧的情况下, 本文也采取了文献[16] 的策略, 即以两个词所有概念进行相互比较, 取概念相似度最大的结果作为其词语相似度的取值。另外, 由于在文本分析中最为有用的是实词, 本实验也以实词为例, 且只考虑同类词性的实词相似度计算。实验结果如表 3 所示。

表 3 词汇相似度计算实验比较 1

词语 1	词语 2	文献[16] 相似度	本文相似度
飞机	战斗机	0.209	0.7
粉红	颜色	0.059	0.4
绿	颜色	0.059	0.273
友谊	特色	0.747	0.455
跑	跳	0.444	0.8

通过对“飞机, 战斗机”“友谊, 特色”“粉红”“绿”“颜色”“跑”“跳”等词语的计算结果可以看到, 本方法有效地解决了义原不对位以及义原所在义原层次树中的深度对概念相似度计算的影响。

表 4 是以文献[16] 的实验数据为例, 得出的更多比较数据, 供大家参考。可以看到, 从其他词语的计算结果来看, 本文也总体与文献[16] 相当。因此, 作为知网概念相似度计算的基本公式, 本方法计算结果更为合理。

表 4 词汇相似度计算实验比较 2

词语 1	词语 2	文献[16]	本文结果
农民	运动员	0.722	0.722
中国	美国	0.936	0.895
粉红	红	1	1
十分	非常	1	1
思考	考虑	1	1
运行	跳舞	0.151	0.308
男人	女人	0.861	0.882

注: 文献[16] 的参数为 $\alpha=1.6$, $\beta_1=0.5$, $\beta_2=0.20$, $\beta_3=1.7$, $\beta_4=1.3$, $\gamma=0.2$, $\delta=0.2$ 。本文计算参数为 $\alpha=1$

4.4 应用实例

概念相似度计算的一个应用实例是 CSCL(计算机支持的协同学习) 的交互分析。交互分析是 CSCL 近年来研究的一个热点, 其主要目的是通过对学生在协作学习中的交互记录进行分析和评价, 监控和反馈交互的状态与质量, 从而促进协作学习中协同知识建构的发生。

衡量交互记录质量的一个重要指标是学生发言与讨论主题的相关性。这样, 可以通过预先设定一些主题词汇, 通过计算学生发言记录中的每个词汇与这些主题词汇的相似度, 得到每个小组成员的发言相关性。相关性计算公式定义为 $relevance = \sum_{i=1}^n [\max_{j=1}^m (Sim(WT_i, WS_j))] / n$ 。其中: WT_i 为第 i 个主题词汇; WS_j 为该成员发言中的第 j 个词汇; $Sim(WT_i, WS_j)$ 为第 i 个主题词汇与第 j 个词汇的相似度。

为验证该方法的有效性, 本文设计并完成了一个关于“超级教师评选”讨论的实验。实验之前, 设计了一个与该话题主题相关的概念图, 并从其节点中抽取出了 23 个主题词汇(图 4)。之后对五个小组共 20 人, 关于此主题的交互记录进行分析, 计算其中主题相关性; 再与人工分析评定的结果进行比较, 计算其一致性。具体数据如表 5 所示。在 SPSS 中计算的 KENDALL 和谐系数为 0.741, 置信度为 $0.08 < 0.1$, 相关度较高。实验证明, 该方法能有效地评估交互言论的主题相关性。

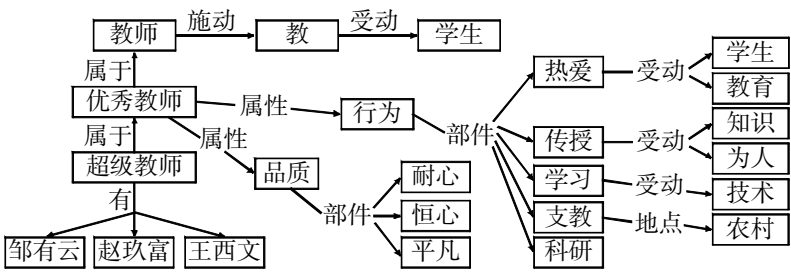


图 4 “超级教师评选”概念图及相关主题词汇

表 5 协作学习中成员发言相关性计算

小组号	成员	人工评定	相关性计算结果
1	A	2	0.70(2)
	B	3	0.69(3)
	C	4	0.72(1)
	D	1	0.67(4)
2	A	1	0.77(1)
	B	2	0.77(1)
	C	3	0.71(3)
	D	4	0.73(4)
3	A	1	0.72(3)
	B	3	0.75(2)
	C	4	0.53(4)
	D	2	0.76(1)
4	A	3	0.74(2)
	B	2	0.74(2)
	C	4	0.71(4)
	D	1	0.75(1)
5	A	4	0.68(4)
	B	2	0.76(2)
	C	1	0.77(1)
	D	3	0.70(3)

5 结束语

形式概念分析是一种对知网进行可视化和表征其概念关

系的有效方法。本文通过对知网进行形式概念分析,得到了一种对知网知识新的表征方法,并在此基础上计算词语的概念相似度。实验证明,该方法对相似度的计算结果比传统方法更为有效。此外,将该方法用于协作学习中交互分析也取得了一定成效。对知网进行形式概念分析还有诸多需要探索和研究的地方,包括如何利用形式概念格让知网在专业领域上进行扩展,如何通过生成的 Hasse 图来对概念进行检索、导航以及推理,再进一步将其应用于专业领域中。这些都将成为笔者未来的研究工作。

参考文献:

[1] ILLER G. WordNet: an on-line lexical database international[J] . Journal of Lexicography, 1990, 3(4) : 265-277.

[2] RICHARDSON S D, DOLAN W B, VANDERVENDE L . MindNet : acquiring and structuring semantic information from text[C] //Proc of COLING-ACL '98. 1998: 1098-1102 .

[3] RICHARDSON S D. Determining similarity and inferring relations in a lexical knowledge base[D] . New York: The City University of New York, 1997.

[4] BAKER C F, FILL M C J, LOWE J B . The Berkeley frame net project[C] //Proc of COLING-ACL '98. 1998: 86-90

[5] 周强, 冯松岩. 构建知网关系的网状表示[J] . 中文信息学报, 2000, 14(6) : 21-27 .

[6] 董振东, 董强. 知网简介[EB/OL] . (2006) . http://www.keen-age.com/zhiwang/c_zhiwang.html.

[7] 董振东. 语义关系的表达和知识系统的建造[J] . 语言文字应用, 1998 (3) : 76-82.

(上接第 31 页)选择参数,而且求解大规模问题比较困难,需要专门的算法和软件。从预测效果看,本文方法并不亚于其他基于氨基酸组成的预测方法,因而本文方法可以作为蛋白质亚细胞定位预测的有效工具。除了氨基酸组成之外,还有许多描述蛋白质序列的方法,可以尝试利用本文的方法进行蛋白质亚细胞定位预测;也可以将本文方法与其他方法相结合,在决策层加以融合进行预测,结果可能会更好。

参考文献:

[1] AKASHIMA H, NISHIKAWA K. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies[J] . Journal of Molecular Biology, 1994, 238(1) : 54-61.

[2] CEDANO J, ALOY P, PEREZ-PONS J A, *et al.* Relation between amino acid composition and cellular location of proteins[J] . Journal of Molecular Biology, 1997, 266(3) : 594-600.

[3] REINHARDT A, HUBBARD T. Using neural networks for prediction of the subcellular location of proteins[J] . Nucleaic Acid Research, 1998, 26(9) : 2230-2236.

[4] CAI Yu-dong, LIU Xiao-jun, CHOU K C. Artificial neural network model for predicting protein subcellular location[J] . Computers and Chemistry, 2002, 26(2) : 179-182.

[5] CHOU K C, ELROD D. Protein subcellular location prediction[J] . Protein Engineering, 1999, 12(2) : 107-118.

[8] 黄伟, 金远平. 形式概念分析在本体构建中的应用[J] . 微机发展, 2005, 15(2) : 28-31.

[9] SANDERSON M, CROFT B. Deriving concept hierarchies from text [EB/OL] . (1999) . <http://dis.shef.ac.uk/mark/cv/publications/pa2pers/my-papers/SIGIR99.pdf>.

[10] GANTER B, WILLE R. Applied lattice theory: formal concept analysis[EB/OL] . (1996) . <http://www.math.tu-dresden.de/~ganter/ps2files/concept.ps>.

[11] BASHIR F I. Ontology construction for structured textual data[EB/OL] . (2003- 04-12) . <http://multimedia.eecs.uic.edu/faisal/downloads/research/cs580-proj6.pdf>.

[12] WOLFF K E. A first course IN formal concept analysis: how to understand line diagrams [EB/OL] . (1993) . <http://www.fbm.fh-darmstadt.de/home/wolff/Publikationen/A-First-Course-in-Formal-Concept-Analysis.pdf>.

[13] LINDIG C. Introduction to concept analysis[EB/OL] . (2002) . <http://www.st.cs.uni-sb.de/~lindig/talks/fca-intro/slides.pdf>.

[14] LINDIG C. Fast Concept Analysis [EB/OL] . (2002) . <http://www.st.cs.uni-sb.de/~lindig/papers/fast-ca/iccs-lindig.pdf>.

[15] 龚永恩, 袁春风, 武港山. 基于语义的词义消歧算法初探[J] . 计算机应用研究, 2006, 23(3) : 41-43, 52.

[16] 刘群, 李素建. 基于知网的词汇语义相似度计算[C] //第三届汉语词汇语义学研讨会. 台北:[出版者不详], 2002.

[17] 许云, 樊孝忠, 张锋. 基于知网的语义相关度计算[J] . 北京理工大学学报, 2005, 25(5) : 411-414.

[18] 易丽萍, 竹勇, 雷小春. 知网在词语相似度计算方面的应用[J] . 人工智能与知识工程, 2005 (1) : 24-26.

[6] HUA Su-jun, SUN Zhi-rong. Support vector machine approach for protein subcellular localization prediction [J] . Bioinformatics, 2001, 17(8) : 721-728.

[7] YUAN Z. Prediction of protein subcellular locations using Markov chain models[J] . FEBS Letters, 1999, 451(1) : 23-26.

[8] FENG Z P, ZHANG C T. Prediction of the subcellular location of prokaryotic proteins based on the hydrophobic index of the amino acids[J] . International Journal of Biological Macromolecules, 2001, 28(3) : 255-261.

[9] CHOU K C. Prediction of protein subcellular attributes using pseudoamino acid composition[J] . Proteins: Structure, Function, and Genetics, 2001, 43(3) : 246-255.

[10] NAKAI K, KANEHISA M. A knowledge base for predicting protein localization sites in eukaryotic cells[J] . Genomics, 1992, 14(4) : 897-911.

[11] NIELSEN H, BRUNAK S, Von HEIJNE G. Machine learning approaches for the prediction of signal peptides and other protein sorting signals[J] . Protein Engineering, 1999, 12(1) : 3-9.

[12] EMANUELSSON O, NIELSEN H, BRUNAK S, *et al.* Predicting subcellular localization of proteins based on their N-terminal amino acid sequence[J] . Journal of Molecular Biology, 2000, 300(4) : 1005-1016.

[13] ANDRADE M A, O 'DONOGHUE S I, ROST B. Adaptation of protein surface to subcellular location[J] . Journal of Molecular Biology, 1998, 276(2) : 517-525.