一种基于FCA的概念相似度算法

林智超,朱国进
(东华大学 计算机学院,上海 201620)

摘 要:概念相似度是计算机自然语言处理研究的重要问题之一。文中描述了两个概念之间相似度计算的一种方法。这种方法是基于形式概念分析属性之上的。相似度的度量由语义相似度和语义距离来定义。首先给出属性相似度,将属性相似度换算成属性距离,接着对属性距离建立网络流最小费用最大流模型得到概念相似度的语义距离,概念相似度的语义距离再换算成最终概念相似度的结果。选择了一个熟悉的领域设计了一个实验,结果表明这种方法是有效的。

关键词:形式概念分析;相似度;属性;网络流
中图分类号:TP301.2 **文献标识码**:A **文章编号**:1673-629X(2008)09-0112-03

A Concept Similarity Algorithm Based on FCA

LIN Zhi-chao, ZHU Guo-jin
(Computer Science College, Donghua University, Shanghai 201620, China)

Abstract: The concept similarity is one of the most important problems in the machine language research. Describes one method to compute the similarity between two concepts. The method is based on the attribute of formal concept analysis. The measurement of concept similarity is defined by semantic similarity and semantic distance. First, get the attribute similarity and transfer to attribute distance. And then build network flow model on attribute distance to get the concept distance which transfer to the final concept similarity. At last uses one example in author's familiar area to show this method is effective for similarity computation.

Key words: formal concept analysis; similarity; attribute; network flow

0 引 言

概念是一种形式化的规范说明[1,2],通俗来说也可以把它看作语言文字中的词语。在信息时代,人们要理解的概念越来越多。勤劳智慧的人民也希望能让机器来理解,从而帮助自己来解决许多问题,提高办事的效率,于是有了计算机自然语言处理研究。概念相似性则是其重要的一个组成部分,也是人工智能应用中一直待以解决的问题。对此,最初由专家完全给定概念相似性直接录入机器。然而这种方法实质上没有摆脱人为地去处理概念,并没有真正让机器去理解概念。后来,有的研究者利用词典中同义词、反义词等组成一个关于词组的树状层次的体系结构来进行计算概念相似性。还有的研究者利用大规模的语料库统计出一个能反映某学科领域的语义相关概念,通过简单的语义关系,计算语义距离来得到相似度。然而这两种方法都依赖于词典、语料库的大小,计算量大,效率不高,并且从广义上讲,不是任何一个概念都可以轻易放进一个特定的层次体系结构中去的。例如,类似于两道程序设计题的相似程度计算,作为一个概念的对象,程序设计题有很多,全部放入特定的概念层次结构有些困难。像这样某领域中随机的对象再组成概念也会有很多,但是用上述的方法根本无法解决相似度判定的问题。基于上述研究现状,文中提出一种通过对形式概念分析(Formal Concept Analysis, FCA)的概念属性研究计算概念相似度的方法。

收稿日期:2007-12-18
基金项目:国家自然科学基金资助项目(60273051)
作者简介:林智超(1983-),男,上海人,硕士研究生,研究方向为计算机语义网络与人工智能;朱国进,博士,副教授,研究方向为计算机语义网络。

1 所用技术基本概念

1.1 形式概念分析基本定义

形式概念分析[3~5]是由德国数学家 Wille 于 1982 年首先提出的,用于概念的发现、排序和显示,所有的概念连同它们之间的泛化、特化关系构成一个概念格。从形式背景中生成概念格的过程实质上是一种概念聚类过程,概念格可用于许多机器学习的任务,目前概念格在信息检索、软件工程和知识发现等方面得到应用。以下是形式概念分析中 4 个很重要的定义:

定义 1:形式背景(Formal - Context) $K = (O, A, R)$ 由集合 O, A 以及它们之间的关系 R 组成,O 中的元素称为对象(Object),A 中的元素称为属性(Attribute)。为了表示一个对象 o 和一个属性 a 在关系 R 中,可以写成 oRa 或 $(o, a) \in R$。

定义 2:给定对象集合 O,对于对象子集 $M \subseteq O$,定义 $M' = \{a \in A \mid \forall o \in A, oRa\}$ 表示 M 中全体对象所共有的属性集。对于属性子集 $N \subseteq A$,定义 $N' = \{o \in O \mid \forall a \in N, oRa\}$ 表示同时具有 N 中所有属性的对象的集合。

定义 3:形式背景 (O, A, R) 中的一个形式概念(Formal - Concept)是一个对 (E, I),其中,$E \subseteq O$,$I \subseteq A$,满足 $E' = I$ 且 $I' = E$,E、I 分别称为形式概念 (E, I) 的外延(Extent)和内涵(Intent)。$\delta(O, A, R)$ 表示形式背景 (O, A, R) 所有形式概念的集合。

定义 4:如果 $(E_1, I_1), (E_2, I_2)$ 是两个形式概念,如果 $E_1 \subseteq E_2$(等同于 $I_1 \subseteq I_2$),那么 (E_1, I_1) 被称为 (E_2, I_2) 的子概念,(E_2, I_2) 被称为 (E_1, I_1) 的超概念,记为 $(E_1, I_1) \leqslant (E_2, I_2)$,关系 $\leqslant$ 称为形式概念之间的序。按此方式有序的 (O, A, R) 所有形式概念的集合被表示为 $\delta(O, A, R)$,并称为形式背景 (O, A, R) 的概念格(Concept - Lattice)。

1.2 语义距离与语义相似度

语言学家认为语义距离和语义相似度之间有着密切的联系[6]。一般来说,两个语义的语义距离越小,语义越相近,相似度越大;反之,两语义距离越大,相似度越小。二者之间可以建立一个简单的对应关系,这种对应关系需要满足以下几个条件:

(1) 两个语义的距离为 0 时,其相似度为 1;

(2) 两个语义的距离为无穷大时,其相似度为 0;

(3) 两个语义的距离越大,其相似度越小(单调下降);

(4) 语义相似度的区间为 0 到 1。

对于两个语义 W_1、W_2,记其相似度 $S(W_1, W_2)$,语义距离为 $D(W_1, W_2)$,可以定义一个满足上述条件的简单关系。

$$S(W_1, W_2) = \frac{\alpha}{D(W_1, W_2) + \alpha} \tag{1}$$

式中,α 是一个可调节参数。

通过转换可得:

$$D(W_1, W_2) = \frac{\alpha}{S(W_1, W_2)} - \alpha \tag{2}$$

用语义距离和语义相似度来定义属性距离、属性相似度、概念距离和概念相似度,即属性相似度为 $S_{\text{Attribute}}(A, B)$,属性距离为 $D_{\text{Attribute}}(A, B)$,概念相似度为 $S_{\text{Concept}}[(E_1, I_1), (E_2, I_2)]$,概念距离为 $D_{\text{Concept}}[(E_1, I_1), (E_2, I_2)]$。

1.3 网络流最小费用最大流

在网络 $D = (V, A, C)$ 中,对应每一条弧 $(V_i, V_j) \in j$,除了已给定的弧 (v_i, v_j) 的容量 $c_{i,j}(c_{i,j} \geqslant 0)$ 外,还给了一个单位流量通过弧 (v_i, v_j) 的费用 $b_{i,j}(b_{i,j} \geqslant 0)$。

f 是 D 的一条可行流,则其总费用为:$b(f) = \sum_{(v_i, v_j) \in A} b_{i,j} f_{i,j}$ 则使得 $b(f)$ 为最小且流量 $v(f)$ 最大的问题称为最小费用最大流问题[7]。最小费用最大流传统求解方法有 Ford - Fulkerson 迭加算法。

2 概念相似度算法的实现

2.1 属性距离与属性相似度的计算

一种最直接的计算属性相似度的方法就是领域专家直接给定。如果某领域的概念的内涵的属性值不是很多的话,这也是非常行之有效的计算属性相似度的方法。文中引用另外一种计算属性相似度的方法以供参考。A、B 为内涵属性,A 和 B 属性相似度计算如下:

$$S_{\text{Attribute}}(A, B) = \frac{P(A, B)}{P(A, B) + P(A, \overline{B}) + P(\overline{A}, B)} \tag{3}$$

通过式(2)转换可以得到 $D_{\text{Attribute}}(A, B)$。

式中,$P(A, B)$ 表示一个对象在某领域集中既具有属性 A 又具有属性 B 的可能性;$P(A, \overline{B})$ 表示一个对象在某领域集中具有属性 A 但不具有属性 B 的可能性;$P(\overline{A}, B)$ 表示一个对象在某领域集中不具有属性 A 但具有属性 B 的可能性。

$P(A, B)$、$P(A, \overline{B})$、$P(\overline{A}, B)$ 计算步骤如下:

对于领域集 E,把它分成具有属性 A 的集合 E^A 和不具有属性 A 的集合 $E^{\overline{A}}$。

对于领域集 E,把它分成具有属性 B 的集合 E^B 和不具有属性 B 的集合 $E^{\overline{B}}$。

对 E^B 进行分类,分成 $E^{A,B}$ 和 $E^{\overline{A},B}$;同样把 $E^{\overline{B}}$ 分成 $E^{A,\overline{B}}$ 和 $E^{\overline{A},\overline{B}}$。

用 $N(E)$ 表示 E 中个数。

$$P(A, B) = \frac{N(E^{A,B})}{N(E)}, P(A, \overline{B}) = \frac{N(E^{A,\overline{B}})}{N(E)},$$

$$P(\overline{A}, B) = \frac{N(E^{\overline{A},B})}{N(E)}$$

2.2 概念距离与概念相似度算法

根据形式概念定义 4,$E_1 \subseteq E_2$(等同于 $I_1 \subseteq I_2$),可以发现两个概念的外延相互之间的关系与内涵相互之间的关系是等价的。可以想象,如果两个概念的属性个数相同,属性值也相同,两个概念是否也非常相似

呢?所以,本节通过属性与概念之间的联系,在上节给出的属性距离与属性相似度计算方法基础上进一步计算概念相似度。计算的方法是基于网络流的最小费用最大流模型。计算方法如下:

(1) 设(E_1, I_1),(E_2, I_2)是一个或多个形式背景的概念,集合 I_1, I_2 的属性基数分别为 n 和 m,提取所有属性

$\{A_i \mid A_i \in I_1, 1 \leqslant i \leqslant n\}$,$\{B_i \mid B_i \in I_2, 1 \leqslant i \leqslant m\}$;

(2) 根据属性相似度计算方法,得出 I_1, I_2 属性相互之间的相似度,记为

$\{S_{\text{Attribute}}(A_i, B_j) \mid 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant m\}$;

(3) 根据式(2)的转换得到 I_1, I_2 属性相互之间的距离,记为$\{D_{\text{Attribute}}(A_i, B_j) \mid 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant m\}$;

(4) 设定一个源点 X,从 X 到$A_i(1 \leqslant i \leqslant n)$引一条有向弧,弧容量分别是集合 I_1 属性的权重 W_i,弧的费用为 1,$(W_1 + W_2 + \cdots + W_n = 1)$;

(5) 设定一个汇点 Y,从每个$B_j(1 \leqslant j \leqslant m)$到 Y 引一条有向弧,弧容量分别是集合 I_2 属性的权重 Q_j,弧的费用为 1,$(Q_1 + Q_2 + \cdots + Q_m = 1)$;

(6) 每个 $A_i(1 \leqslant i \leqslant n)$ 向 $B_j(1 \leqslant j \leqslant m)$ 引一条有向弧,弧容量都是 1,费用为 $D_{\text{Attribute}}(A_i, B_j)$;

(7) 利用网络流最小费用最大流算法对从源点 X 到汇点 Y 的最小费用值进行计算得到 Z(最大流恒定为 1);

(8)(E_1, I_1), (E_2, I_2) 的概念距离 $D_{\text{Concept}}[(E_1, I_1), (E_2, I_2)]$是 $Z - n - m$ (概念距离仅包括 $A_i(1 \leqslant i \leqslant n)$ 到 $B_j(1 \leqslant j \leqslant m)$ 的费用值);

(9) 通过式(1) 可以求得 $S_{\text{Concept}}[(E_1, I_1), (E_2, I_2)]$。

其计算流程图如图 1 所示。

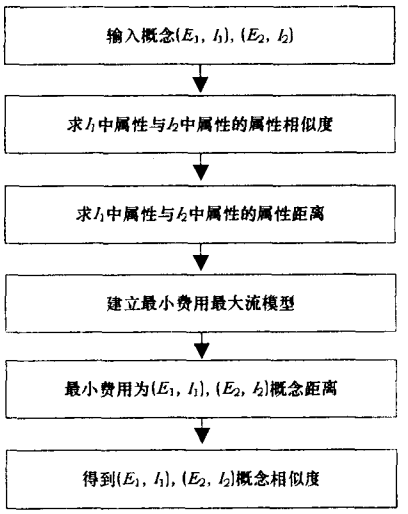

图 1 相似度计算流程图

3 实验结果与分析

为了验证文中提出的概念相似度的算法的可行性,设计一个笔者熟悉的形式背景做实验。笔者多年参与 ACM 竞赛,ACM 竞赛是一个考核选手算法和数据结构以及程序设计语言编写能力的竞赛,每道题目的解题方法之间也有许多相似之处值得研究。先提取国际著名的 Valladolid Online Judge Site 题库网站(http://online-judge.uva.es/problemset/)上的几道题目,拿出来做相似度分析。

选取网站上 P374、P495、P10229、P10334 四道题目,针对这四道题目建立如表 1 所示的形式背景(首行表示属性集,首列表示对象集,X 表示对象拥有此属性,空白表示对象没有此属性)。

表 1 形式背景表

	(a)高精度	(b) Fibonacci	(c) Big Mod
(1)P374			X
(2)P495	X	X	
(3)P10229		X	X
(4)P10334	X	X	

然后选取几个不同的选手,通过自己拟定属性集的权重和式(1)、式(2)中 α 可调节参数,通过两种方法对以下 4 组概念进行相似度的计算取平均值(见表 2)。方法 1 是文中提出的基于属性的概念相似度算法,方法 2 是人为进行判断。

表 2 实验结果表

序号	概念 1	概念 2	方法 1	方法 2
1	[(2), (a, b)]	[(4), (a, b)]	1.00	0.95
2	[(4), (a, b, c)]	[(1), (a, b, c)]	0.23	0.10
3	[(2), (a, b, c)]	[(3), (a, b, c)]	0.48	0.55
4	[(3,4), (a, b, c)]	[(1,2), (a, b, c)]	0.84	0.90

从实验的结果来看,与人为的判断较为吻合,即文中的方法是有效的。再从实验结果来分析,方法 1 与方法 2 的结果有误差,这些误差主要来自于人为判断对概念属性值之间的相似度的估计不足。故属性相似度对文中提出的算法的结果的重要性可见一斑。

再从实验序号 1,可以看到,因为两个概念十分相似,而两个概念分别只对应着(P495,P10334)两道题目,意味着一选手如果解决了 P495 题,应该很快就可以解决 P10334,反之亦然。从实验序号 2 来看,两个概念对应着(P10334,P374)两道题目,且相似度不大,意味着 P10334 题的解题方法对 P374 题的帮助不大。从实验序号 4,又可以看出如果解决了 P10229,P10334

(下转第 126 页)

提供较高的安全性能,但协议为 Funk 公司专有,用户需为请求者和认证服务器软件付费,增加了无线局域网的部署成本。LEAP 以前是 Cisco 专有协议,仅与 Cisco 无线适配器一起使用,虽然后来其它制造商经授权也可使用 LEAP,但即使在 LEAP 认证时执行强密码策略,也无法满足高安全要求的无线局域网部署。在上述各 EAP 类型中,TLS 安全性最高,TLS 是一个 IETF 标准,是无线客户端和 RADIUS 服务器上最受支持的一个标准,但 TLS 不仅需要在服务器端安装证书,而且在各个无线工作站上也要安装客户端证书,增加了无线局域网的部署难度,TLS 适合于对安全性有较高要求的大型无线局域网部署中。PEAP 是一种基于安全密码的认证协议,得到了 Microsoft、Cisco 和 RSA Security 等软、硬件厂商的广泛支持,PEAP 可以兼容几乎全部厂商的全部设备,拥有了相当规模的市场占有率;由于 PEAP 使用服务器单边证书来认证无线局域网客户端,且 PEAP 与 Windows 操作系统的良好协调性,以及可以通过 Windows 组策略进行管理的特性,从而简化了安全无线局域网的部署和管理。

4 结束语

802.11i 标准中,采用 802.1x 结合 EAP 并选择高级加密标准 AES,可以为无线局域网提供强健的安全保障。具体应用哪一类 EAP,要综合考虑协议的安全性、通用性以及能否更方便部署等因素。由于 EAP-TLS 和 PEAP 自身的特点,使其成为部署无线局域网时优先考虑的 EAP 类型。Windows Server 2003、Windows XP 以及 Windows Vista 都提供了对 EAP-TLS 和 PEAP 的内置支持,Microsoft 公司提供了两套无线局域网安全解决方案[7]:使用 EAP-TLS 的"确保无线 LAN 的安全-证书服务解决方案"和"使用 PEAP 和密码确保无线 LAN 的安全",分别用于信息技术环境相对比较复杂的大型机构和中、小型企业部署高安全性的无线局域网。

参考文献:

[1] IEEE. IEEE standard for local and metropolitan area networks-Port-Based Network Access Control[S]. USA:[s.n.], 2001.

[2] 袁建国,方宁生,姜 浩. 802.1x:基于端口的访问控制协议[J]. 微机发展(现名:计算机技术与发展),2005,15(12):160-161.

[3] 周 晓,王芙蓉,郭 毅. 802.1x 在分布式防火墙中的应用[J]. 计算机技术与发展,2006,16(12):245-246.

[4] 楼颖明,罗汉文. 802.11 无线局域网的安全方案分析[J]. 通信技术,2002(9):79-80.

[5] Microsoft. 选择无线 LAN 的安全策略[EB/OL]. 2004-05-27. http://www.microsoft.com/china/technet/security/guidance/peap-int.mspx.

[6] Intel. 无线安全—802.1x 和 EAP 类型[EB/OL]. 2006-10-18. http://www.intel.com/support/cn/wireless/wlan/sb/cs-008413.htm.

[7] Microsoft. 使用 PEAP 和密码确保无线 LAN 的安全[EB/OL]. 2004-05-27. http://www.microsoft.com/china/technet/security/guidance/peap-1.mspx.

(上接第 114 页)

两题应该很快也能解决 P374,P495 两题。由此可见,概念相似度在对选手训练程序设计也是有帮助的。

4 结束语

概念相似度的计算除了在上一节描述以外,还有许多领域有着广泛的应用,例如信息检索,它是机器理解概念进行推理的重要一步。在国内外相关研究基础上,通过 FCA 的概念,根据笔者多年对程序设计算法的研究,巧妙地结合了网络流算法,提出基于属性的概念相似度的算法,这也是文中创新之处。通过实验可以看出,所提出的算法是可行的。当然该算法也有进一步挖掘的地方,比如与语义距离三角形原理结合使最终结果更人性化、更贴合实际情况。

参考文献:

[1] Borst W N. Construction of Engineering Ontologies for Knowledge Sharing and Reuse [D]. Enschede: University of Twente, 1997.

[2] Studer R, Benjamins V R, Fensel D, et al. Knowledge engineering, principles and methods[J]. Data and Knowledge Engineering, 1998, 25(1-2):161-197.

[3] Ganter B, Wille R. Formal Concept Analysis: Mathematical Foundations[M]. Heidelberg: Springer, 1999.

[4] Wille R. Restructuring lattice theory: An approach based on hierarchies of concepts[C]//In: Rival I. Oredered sets. Dordrecht-Boston: Reidel, 1982: 445-470.

[5] Ivkovic I, Kontogiannis K. Towards Automatic Establishment of Model Dependencies Using Formal Concept Analysis[J]. International Journal of Software Engineering and Knowledge Engineering (IJSEKE), 2006, 16(4): 499-522.

[6] 刘 群,李素建. 基于《知网》的词汇语义相似度计算[J]. Computaional Linguistics Chinese Language Processing, 2002, 7(2): 59-76.

[7] 胡运权,郭耀煌. 运筹学教程[M]. 第 3 版. 北京:清华大学出版社,2007.