基于限界传递相似度图的FCA概念相似度计算方法

黄宏涛[1] 吴忠良[1] 万庆生[2] 黄少滨[2]
(河南师范大学河南省高校教育信息工程技术研究中心 新乡 453007)[1]
(哈尔滨工程大学计算机科学与技术学院 哈尔滨 150001)[2]

摘 要 使用相似度图计算FCA概念相似度需要构造相似关系的传递闭包,对于复杂问题会导致相似度图规模过大,从而影响相似度评价的效率。为了降低相似度图规模,提出一种基于限界传递相似度图的FCA概念相似度计算方法。该方法首先通过限定传递相似关系的长度来避免构造相似关系的传递闭包,得到的限界传递相似度图中忽略了长度超过界限且对区分FCA概念无用的传递相似关系,能够有效压缩相似度图的规模;然后给出了动态传递相似度计算方法和由限界传递相似度图构建二部图的方法。实验结果表明,使用限界传递相似度图能够在不损失计算结果准确度的情况下有效提高FCA概念相似度计算的效率。

关键词 FCA概念相似度,相似度图,传递相似关系,限界传递
中图法分类号 TP391.1 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.1.063

FCA Concept Similarity Computation Based on Bounded Transitive Similarity Graph

HUANG Hong-tao[1] WU Zhong-liang[1] WAN Qing-sheng[2] HUANG Shao-bin[2]
(Engineering Research Center of Henan Provincial Universities for Education Information, Henan Normal University, Xinxiang 453007, China)[1]
(College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China)[2]

Abstract It is necessary to construct the transitive closure of similarity relation in the case of computing similarity between FCA concepts by means of similarity graph. This method will lead to large scale similarity graph for complex problem, which may affect the efficiency of similarity evaluation. A bounded transitive similarity graph based FCA concept similarity computing method was proposed in order to reduce the size of similarity graph. This method can avoid constructing the transitive closure of similarity relation by adding a bound on transitive similarity relation, and the bounded transitive similarity graph obtained does not contain the transitive relation whose length beyonds the bound, and this omitted transitive relation is useless to distinguish different FCA concepts, which makes it possible to compress the scale of similarity graph. Then a dynamic transitive similarity computation method and a bipartite graph construction method using bounded transitive similarity graph were given. Experimental results show that this bounded transitive similarity graph based method improves the efficiency of FCA concept computation effectively without the loss of accuracy.

Keywords FCA concept similarity, Similarity graph, Transitive similarity relation, Bounded transitivity

1 引言

形式概念分析(Formal Concept Analysis,FCA)是一种重要的数据分析工具[1],目前是人工智能领域的研究热点之一。FCA提供了一个对数据进行构建、分析和可视化的概念框架,它的目的是使构建的概念体系更加可信,为用户分析和组织领域知识提供方法支持。FCA的理论基础是概念格,它在本体工程[2-5]、软件工程[6]、知识发现[7]、语义检索[8]等领域中有着广泛应用。这些应用中的核心环节之一是FCA概念相似度的评价。

FCA应用范围的不断扩大进一步促进了FCA概念相似度评价方法的发展,使得对FCA概念相似度评价的需求不断增强。文献[9]提出了一种使用相似度图对FCA概念相似度进行评价的方法,该方法在用于符号本体开发的SymOntos项目中得到了成功应用,相似度图的引入提高了FCA概念相似度评价结果的准确性。文献[10]提出了一种本体支持下的FCA概念相似度评价方法,该方法也使用了文献[9]提出的相似度图来保证获得较高准确率的评价结果。然而,由于相

到稿日期:2014-02-07 返修日期:2014-04-11 本文受国家科技支撑计划项目(2012BAH08B02),河南省科技攻关项目(082400420250,112300410008),河南省教育厅科学技术研究重点项目(13A520508),河南师范大学博士科研启动基金项目(qd12107),青年科学基金项目(2013qk33)资助。

黄宏涛(1980-),男,博士,副教授,主要研究方向为问答系统、模型检测,E-mail:huanght@outlook.com;吴忠良(1975-),男,硕士,副教授,主要研究方向为数据挖掘;万庆生(1985-),男,博士生,主要研究方向为问答系统、数据挖掘;黄少滨(1975-),男,博士,教授,主要研究方向为数据挖掘、模型检测。

似度图是相似关系的自反、对称、传递闭包，传递闭包会导致相似度图规模随领域问题规模的增长急剧膨胀，从而影响相似度计算的效率。

传递相似关系长度的增加会使传递相似度值无限趋近于零。因此，对于给定规模的传递相似度图，长度超过一定数值的传递相似关系对于区分 FCA 概念是没有意义的。受文献[11]的启发，本文提出一种基于限界传递相似度图的 FCA 概念相似度计算方法，其基本思想是根据问题规模，为相似度图中的传递相似关系设定长度界限，合理的长度界限可以在保证计算结果准确性的前提下有效缩减相似度图的规模，提高算法效率。

2 相关概念

一个领域本体包含由一组相互关联的概念构成的集合，该集合中的每个概念都是一个形式化定义，它是给定领域中的一个无二义性的概念。由于相似度图是由领域本体中的相似关系构成，因此下面给出的本体定义仅关注实体间的相似关系。

定义 1 领域本体 O 是一个二元组 (E_O, Sim)，E_O 为实体名称集合，Sim 为语义相似关系集合。

一般来说，语义关系可以是泛化关系(ISA)、部分关系、相关关系以及相似关系等。为了利用本体中定义的相似度关系进行 FCA 概念相似度计算，定义 1 把语义关系限定在相似关系上。语义相似度关系 Sim 可以使用一个三元组来定义：

$$Sim(c_i, c_j, as(c_i, c_j)) \tag{1}$$

其中，c_i、c_j 为实体名称，$as(c_i, c_j)$ 是一个小数，其数据域为 $[0.0-1.0]$，表示由本体 O 得出的实体 c_i 和 c_j 之间的公理相似度。这里的相似度是由特定领域专家通过共识系统建立的。由定义可知，领域本体 O 是对实体名称和语义相似度集合的规约。

定义 2 形式背景为一个三元组 $C=(U,A,R)$，它由对象集合 U、属性集合 A 以及关系 R 构成，其中 R 是 U 和 A 上的二元关系，称 U 为形式对象，A 为形式属性，$(u,a)\in R$ 或 uRa 表示对象 u 具有属性 a。

3 FCA 概念相似度计算

3.1 限界传递相似度图

对给定的领域本体和形式背景，可以由本体中的实体和形式背景中的属性构建相似度图。文献[9]给出的相似度图是相似关系的自反、对称和传递闭包，在给出限界传递相似度图的概念前，下面先给出相似度图的定义，不同的是，该定义不考虑传递相似关系。

定义 3 给定领域本体 $O(E_O, Sim)$ 和形式背景 (U,A,R)，令 $\xi=(E_O\cup A)$，$\Gamma_{(O,A)}$ 为一个相似度三元组 $\langle c_i, c_j, as(c_i, c_j)\rangle$ 集合，其中 $c_i\in\xi, c_j\in\xi$，称 $\Gamma_{(O,A)}$ 为 O 和 (U,A,R) 的相似度图当且仅当下列条件成立：

- 对任意 $c_i\in\xi、c_j\in\xi$，$(c_i, c_j, as(c_i, c_j))\in Sim$ 蕴含 $\langle c_i, c_j, as(c_i, c_j)\rangle\in\Gamma_{(O,A)}$；
- 对任意 $c\in\xi$，都有 $\langle c, c, as(c,c)\rangle\in\Gamma_{(O,A)}$ 成立，且 $as(c, c)=1.0$；
- 对任意 $c_i\in\xi$ 和 $c_j\in\xi$，有 $as(c_i, c_j)=as(c_j, c_i)$ 和 $\langle c_i, c_j, as(c_i, c_j)\rangle\equiv\langle c_j, c_i, as(c_j, c_i)\rangle$ 成立。

由上述定义可知，在相似度图 $\Gamma_{(O,A)}$ 上，可以把 $as(c_i, c_j)$ 视为 c_i 和 c_j 的公理相似度。例如，对于文献[10]中给出的形式背景 (U,A,R)，表 1 给出了其二维关系。

表 1 形式背景的二维关系表

	a	b	c	d	e
1		√	√	√	
2			√	√	√
3	√		√	√	√
4	√				√

于是可得 (U,A,R) 对应的概念格如图 1 所示。

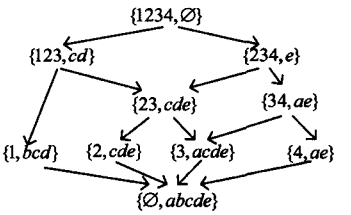

图 1 形式背景对应的概念格

对于与 (U,A,R) 相关的领域本体 $O(E_O, Sim)$，其中

$E_O=\{a,a',b,b',d,d'\}$

$Sim=\{(a,a',0.8),(b',b,0.9),(d,d',0.7),(b',d',0.2),(a,b,0.1)\}$

由于对领域术语使用习惯的不同，用户表述相近意义的概念时，使用的术语可能会有较大的差别。例如“职工”和“员工”，“医疗保险”和“医保”。使用相似度图能够有效地计算出 FCA 概念间的相似程度。令 $\Gamma_{(O,A)}$ 为 $O(E_O, Sim)$ 和 (U,A,R) 的相似度图，由定义 3 可得图 2 所示的相似度图 $\Gamma_{(O,A)}$。

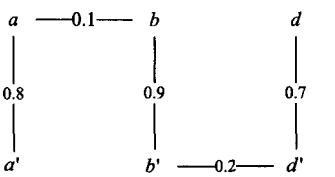

图 2 $O(E_O, Sim)$ 和 (U,A,R) 的相似度图

由定义 3 得出的相似度图是相似关系的自反和对称闭包(图 2 中略去了自反关系，两点间的连线表示两点间的对称关系)，这种定义方法通过忽略相似关系的传递性来达到缩小相似度图规模的目的，这样既可以在 FCA 相似度匹配过程中引入本体支持，同时又不会因相似度图规模过大而影响计算效率。但是完全不考虑相似关系的传递性又会忽略一些潜在的相似关系，从而导致相似度匹配的精度下降。例如，设“员工”和“职工”之间的相似度为 0.8，“职工”和“职员”之间的相似度为 0.9，则“员工”和“职员”之间也存在一定程度的相似性，此处使用两个相似关系相似度之积作为传递关系的相似度，即“员工”和“职员”之间的相似度为 0.72。如果在图 2 基础上增加已有相似关系的传递关系，即计算图 2 的传递闭包，则所得相似度图的规模会较为庞大，从而影响计算效率。为了平衡效率和精度之间的关系，下面给出限界传递相似度图的定义。

定义 4 设 $\Gamma_{(O,A)}$ 为 $O(E_O, Sim)$ 和 (U,A,R) 的相似度图，对于自然数 η 和 κ，有

$\langle c_1, c_2, as(c_1, c_2)\rangle\in\Gamma_{(O,A)}$

$\langle c_2, c_3, as(c_2, c_3)\rangle\in\Gamma_{(O,A)}$

…

$\langle c_{\kappa-1}, c_\kappa, as(c_{\kappa-1}, c_\kappa)\rangle \in \Gamma_{(O,A)}$

蕴含

$$\bigcup_{i=3}^{\kappa} \langle c_1, c_i, \prod_{j=1}^{i-1} as(c_j, c_{j+1})\rangle \subseteq \Gamma_{(O,A)}$$

仅在 $3 \leqslant \kappa \leqslant \eta$ 时成立，则称 $\Gamma_{(O,A)}$ 为限界传递相似度图，记为 $\Gamma^{\eta}_{(O,A)}$，称 $\Gamma^{\eta}_{(O,A)}$ 中的传递相似关系为 η-传递相似关系。

由定义 4 知，当 $\eta=3$ 时，图 2 所示相似度图 $\Gamma_{(O,A)}$ 的限界传递相似度图 $\Gamma^{3}_{(O,A)}$ 如图 3 所示。

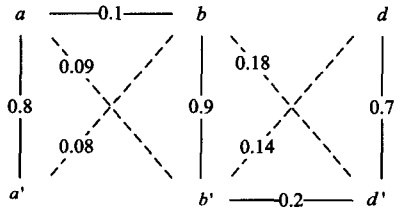

图 3 $O(E_O, Sim)$和(U, A, R)的限界传递相似度图 $\Gamma^{3}_{(O,A)}$

3.2 限界传递相似度计算

注意到随着传递关系的传递长度的增加，传递相似度值就会无限趋近于 0。也就是说传递次数较多时所得到的传递相似度值可以忽略不计。因此，选择适当长度的 η 值既可以使相似度图保持较小的规模，又不至影响到相似度计算结果的精确度。

在实际应用中，并不直接把 η-传递相似关系存储在限界传递相似度图中，限界传递相似度图中的相似关系是动态添加的，也就是说限界传递相似度图初始时不包含传递相似关系，传递相似关系是在使用过程动态添加的。在这种情况下进行 FCA 概念相似度计算时，如果传递相似关系在限界传递相似度图中存在，则直接从相似度图中获取；否则就需要使用 On-The-Fly 方法动态计算限界传递相似关系。下面给出动态计算限界传递相似关系的算法(Dynamic η-Bounded Transitivity Similarity Graph，Dη-TSG)。

算法 1 Dη-BTSG 算法

输入：$\Gamma^{\eta}_{(O,A)}$ 上的顶点 v_1、v_2

输出：若 v_1、v_2 存在 η-相似关系，返回 η-相似度，否则返回 0.0

```
1. If ⟨v1, v2, as(v1, v2)⟩∈Γ^η_(O,A) Then
2.   Return as(v1, v2);
3. Else
4.   令 S_η := ∅; i := 1; Insert(S_η, v1); {S_η 为容量为 η 的堆栈}
5.   While !IsEmpty(S_η) do
6.     v_top := GetTop(S_η);
7.     If v_top = v2 then
8.       令⟨v1, v2, ∏_{j=1}^{|S_η|-1} as(S_η[j], S_η[j+1])⟩∈Γ^η_(O,A);
9.       Return ∏_{j=1}^{|S_η|-1} as(S_η[j], S_η[j+1]);
10.    Else If ∃v∉S_η && ⟨v_top, v, as(v_top, v)⟩∈Γ^η_(O,A) then
11.      If !IsFull(S_η) then
12.        Put(S_η, v);
13.      Else
14.        Pop(S_η);
15.      EndIf
16.  Else
17.      Pop(S_η);
18.  EndIf
19.    EndWhile
20.  If IsEmpty(S_η) then
21.    Return 0.0;
22.  EndIf
23. EndIf
```

当需要计算相似度的两个顶点在 $\Gamma^{\eta}_{(O,A)}$ 上存在相似关系时，Dη-TS 算法直接使用相似度图 $\Gamma^{\eta}_{(O,A)}$ 上的相似度(1－3 行)；否则，就需要动态计算两顶点间的 η-传递相似度：算法从顶点 v_1 开始寻找是否存在从 v_1 到 v_2 的长度为 η 的路径(4－23 行)，如果找到则把该传递相似关系加入 $\Gamma^{\eta}_{(O,A)}$，同时返回 v_1 到 v_2 的传递相似度(7－11 行)，如果找不到就返回 0.0 作为 v_1 和 v_2 的相似度。η-传递相似度使用深度优先搜索方法展开从顶点 v_1 出发到 v_2 的路径，为了保证搜索的深度为 η，第 4 行定义了一个容量为 η 的堆栈 S_η，S_η 的容量确保了搜索到的从 v_1 到 v_2 的路径长度一定小于等于 η(11－13 行)，从而避免了无限制的搜索，减小了搜索状态空间的规模；当搜索过程达到搜索深度时还没有找到从 v_1 到 v_2 的路径，则算法将栈顶元素弹出，并从上一个结点开始继续展开深度优先搜索(13－15 行)。

使用 Dη－TS 算法无需在构建 $\Gamma^{\eta}_{(O,A)}$ 时预先初始化好所有的限界传递相似关系，传递相似关系是在使用的同时动态构建的。Dη-TS 算法限制了搜索传递相似度的路径长度，能够有效降低在相似度图上计算 FCA 概念相似度的时间开销。

下面以表 1 给出的形式背景和领域本体 $O(E_O, Sim)$ 为例介绍使用动态限界传递相似度图计算 FCA 概念相似度的过程。使用文献[10]给出的方法计算图 1 所示概念格中两相邻概念((2,3),(c,d,e))和((3,4),(a,e))的相似度，取 $w=0.5$，它们的相似度计算可使用限界传递相似度图进行。由定义 3 知⟨(e,e,1.0)⟩，a 和 d 之间没有直接的相似度定义，需要计算其 η-传递相似度。如果令 $\eta=5$，则有⟨(a,d,0.0126)⟩，由于 $r=2, m=3$，因此：

$$Sim[((2,3),(c,d,e)),((3,4),(a,e))]$$
$$=\frac{1}{2}\times\frac{1}{2}+\frac{1}{3}\times(1.0+0.0126)\times(1-\frac{1}{2})$$
$$=0.419$$

概念格中概念节点和它的直接后继节点间的相似度较大。对于概念((2,3,4),(e))，它与其后继概念((3,4),(a,e))间的相似度计算如下：

$$Sim[((2,3,4),(e)),((3,4),(a,e))]$$
$$=\frac{2}{3}\times\frac{1}{2}+\frac{1}{2}\times(1.0)\times(1-\frac{1}{2})$$
$$=0.583$$

如果两概念不直接相关，则它们间的相似度较低。例如，概念((1,2,3),(c,d))和概念((3,4),(a,e))间的相似度计算如下：

$$Sim[((1,2,3),(c,d)),((3,4),(a,e))]$$
$$=\frac{1}{3}\times\frac{1}{2}+\frac{1}{2}\times(0.0126+0.0)\times(1-\frac{1}{2})$$
$$=0.170$$

从上述实例来看，在限界传递相似度图支持下计算出的相似度能够精确地区分不同的 FCA 概念。

3.3 二部图构造

在 FCA 概念相似度计算过程中，对算法性能影响最大的

步骤是寻找偶对权值之和最大的候选集，该步骤需要对两个概念所有可能的候选集中的元素权值进行穷尽计算。把这个问题转换为求解二部图最大权匹配问题能够有效降低算法复杂度[12]。因此，实际应用中需要把限界传递相似度图转换为二部图。下面先给出二部图的定义，然后给出由限界传递相似度图构造带权二部图的方法。

定义 5 对于无向图 $G(V,E)$，V 为 G 的顶点集，E 为 G 的边集，如果令 $V=V_1 \cup V_2$，且有 $V_1 \cap V_2=\varnothing$，即 V_1、V_2 是 V 的一个划分，若 E 中任意一条边的两个端点分别来自 V_1、V_2，则称 $G_b(V_1,V_2,E)$ 为一个二部图。

对于两个概念 (E_1,I_1) 和 (E_2,I_2)，为了找出它们所有候选集中偶对权值最大的一个，需要使用限界传递相似度图 $\Gamma^{\eta}_{(O,A_i \cup A_j)}$ 构造带权二部图，A_i、A_j 分别为这两个概念所属形式背景的属性集。下面给出在限界传递相似度图支持下构造带权二部图的方法。

对由 I_1 和 I_2 中元素构成的偶对 $\langle a_i,b_i\rangle$，$a_i \in I_1$，$b_i \in I_2$，带权二部图 $G_b(A_i,A_j,E)$ 的构造需要经过以下步骤：

1) $Sim(a_i,b_i,as(a_i,b_i))$ 在限界传递相似度图 $\Gamma^{\eta}_{(O,A_i \cup A_j)}$ 上有定义，则令 $(a_i,b_i) \in E$，其中 A_i 是概念 (E_1,I_1) 所属形式背景的属性集，A_j 是概念 (E_2,I_2) 所属形式背景的属性集，且 (E_1,I_1) 和 (E_2,I_2) 所属的形式背景可以相同，也可以不同。

2) 假设 A_i、A_j 中的元素各不相同（如果 I_1 和 I_2 中都包含属性 a，则认为这两个 a 是不同的属性，分别表示为 a_i 和 a_j），则 A_i、A_j 和 E 构成一个二部图，令其为 $G_b(A_i,A_j,E)$。

3) 如果把 $\Gamma_{(O,A_i \cup A_j)}$ 上的公理相似度值 $as(a_i,b_i)$ 作为 E 中边的权值，则 $G_b(A_i,A_j,E)$ 成为一个带权二部图。

4 实验结果与分析

本节使用社保中五险问题集对限界传递相似度图支持下的 FCA 概念相似度计算方法进行实验，实验目的是验证限界传递相似度图对 FCA 概念相似度计算准确率与时间性能的影响。实验使用基于 FCA 的问句相似度计算方法（FCA based Question Sentence Similarity，FCAQSS）对社保问题集中的 35 个典型问句进行相似度计算，其核心步骤是 FCA 概念间的相似度计算，计算结果的准确率、时间性能能够直接反映 FCA 概念相似度计算结果的准确率和时间性能。实验分为两个部分：首先对使用传递闭包相似度图（Transitive Closure Similarity Graph，TCSG）和使用限界传递相似度图（Bounded Transitive Similarity Graph，BTSG）时 FCAQSS 算法的准确率进行对比；然后对两种情况下 FCAQSS 算法的时间性能进行分析。

图 4 给出了使用传递闭包相似度图和限界传递相似度图时 FCAQSS 计算结果的准确率。其中使用限界传递相似度图时，取 η 为 3、7、11（D3-BTSG、D7-BTSG、D11-BTSG）3 种情况进行实验。实验结果列出了 FCAQSS 算法在不同噪声比下对 35 个问句进行相似度计算时的准确率均值，目的是分析两种相似度图对 FCA 概念相似度匹配准确率的影响。从实验结果可以看出，使用 D3-BTSG 时的准确率不如 TCSG，而且其差距较大，造成这种情况的主要原因是 D3-BTSG 只动态计算长度为 3 的传递相似度，忽略了所有长度超过 3 的传递相似关系。D7-BTSG 要明显好于 D3-BTSG，但仍不如 TCSG，这说明长度为 4 到 7 的传递相似关系对计算结果准确率仍有显著影响；但两者之间的差距已经不大，这意味着长度超过 7 的传递相似关系对算法准确率的影响已经较小。D11-BTSG 的准确率与 TCSG 不相上下，甚至出现局部等于或优于 TQSS 的情况，这说明在本次实验对象下使用长度为 11 的传递相似关系计算出的 FCA 概念相似度已经接近最合理的情况。总的来说，传递相似关系较长时其可信程度也在下降，对传递相似关系进行合适的限界能够取得较高的准确率，但是 η 值需要根据相似度图的规模以及传递相似关系的复杂程度来确定。

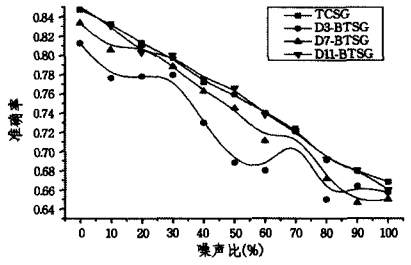

图 4 TCSG 和 BTSG 对 FCAQSS 准确率的影响

图 5 给出了使用传递闭包相似度图和限界传递相似度图时 FCAQSS 算法的时间性能。实验结果列出了对其中 20 个问句进行相似度计算时的平时时间代价。总的来看，D3-BTSG、D7-BTSG、D11-BTSG 的时间性能都明显优于 TCSG，原因是 BTSG 限制了传递相似关系的长度，使得 FCA 概念相似度计算能够在较短的时间获取有效的传递相似度，而 TCSG 没有限制传递相似关系的长度，这使得 FCA 概念相似度计算的时间代价较高。注意到 D3-BTSG 的时间性能略优于 D7-BTSG 和 D11-BTSG，而 D7-BTSG 的时间性能也略好于 D11-BTSG，但是三者的性能差异并不大，其主要原因是它们之间的区别在于动态计算传递相似度时对传递相似关系长度的限制不同，而 BTSG 算法的搜索深度直接和 η 相关，η 越小时算法的时间性能越好，而在本节实验环境下 D3-BTSG、D7-BTSG 和 D11-BTSG 的 η 值差异并不悬殊，所以三者的时间性能差异并不明显。

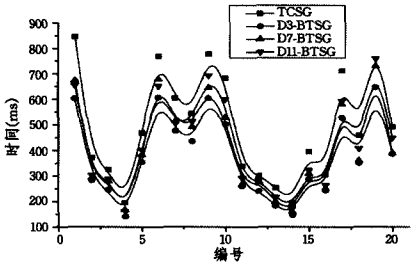

图 5 TCSG 和 BTSG 对 FCAQSS 效率的影响

结束语 限界传递相似度图通过为传递相似关系的长度设定界限，忽略了长度较大的传递相似关系，这些长度超过一定数值的传递相似关系对于区分 FCA 概念的作用可以忽略不计，但是会消耗大量的计算。实验结果表明通过构造传递相似度图可以在保证计算结果准确性的前提下缩小相似度的规模，避免额外的无用计算，从而提高 FCA 概念相似度的计算效率，进而进一步提高基于 FCA 的问句相似度计算方法的实用性。然而，如果算法能够根据具体问题的规模自动设定传递相似关系的界限，就可以有效改善算法的易用性，文献[13-15]等工作提出的动态传递闭包方法为解决这一问题提

供了思路，进一步工作将重点研究限界传递相似度图界限的自适应调整方法。

参 考 文 献

[1] Wille R, Yahia S, Nguifo E, et al. Formal Concept Analysis as Applied Lattice Theory[M]// Concept Lattices and Their Applications. Springer Berlin Heidelberg, 2008, 4923: 42-67

[2] Ning L, Guanyu L, Li S. Using formal concept analysis for maritime ontology building[C]// 2010 International Forum on Information Technology and Applications (IFITA). 2010, 2: 159-162

[3] Sánchez D, Batet M, Isern D, et al. Ontology-based semantic similarity: A new feature-based approach[J]. Expert Systems with Applications, 2012, 39(9): 7718-7728

[4] Li D, Du J, Yao S. Research on Computer Science Domain Ontology Construction and Information Retrieval[J]. Knowledge Engineering and Management, 2011, 123: 603-608

[5] Kang X, Li D, Wang S. Research on domain ontology in different granulations based on concept lattice[J]. Knowledge-Based Systems, 2012, 27: 152-161

[6] Maarek Y S, Berry D M, Kaiser G E. An information retrieval approach for automatically constructing software libraries[J]. IEEE Transactions on Software Engineering, 1991, 17(8): 800-813

[7] Kumar C A. Knowledge discovery in data using formal concept analysis and random projections[J]. International Journal of Applied Mathematics and Computer Science, 2011, 21(4): 745-756

[8] Formica A. Semantic web search based on rough sets and fuzzy formal concept analysis[J]. Knowledge-Based Systems, 2012, 26: 40-47

[9] Formica A, Missikoff M. Concept similarity in SymOntos: an enterprise ontology management tool[J]. The Computer Journal, 2002, 45(6): 583-594

[10] Formica A. Ontology-based concept similarity in formal concept analysis[J]. Information Sciences, 2006, 176(18): 2624-2641

[11] Biere A, Cimatti A, Clarke E M, et al. Bounded model checking [J]. Advances in computers, 2003, 58: 117-148

[12] Bayati M, Shah D, Sharma M. Maximum weight matching via max-product belief propagation[C]// Proceedings of the 2005 IEEE International Symposium of Information Theory. 2008, 54: 1763-1767

[13] Yellin D M. Speeding up dynamic transitive closure for bounded degree graphs[J]. Acta Informatica, 1993, 30(4): 369-384

[14] Sankowski P, Mucha M. Fast dynamic transitive closure lookahead[J]. Algorithmica, 2010, 56(2): 180-197

[15] 舒虎，崇志宏，倪巍伟，等. X-Hop：传递闭包的多跳数压缩存储和快速可达性查询[J]. 计算机科学，2012(3)：144-148

(上接第 256 页)

结束语 本文提出的自适应模型引入了上游路段速度、下游路段最新速度、下游路段最新花时、时间段和路况拥挤程度等动态信息作为模型特征，其性能优于静态预测模型和动态预测模型。而且自适应预测模型是基于日期、时间段和站点波动相似性进行路段分段组合预测的，在保证预测准确性的基础上，提高了预测效率。但是本文提出的预测模型仍然有许多需要改进的地方，如新特征的加入对系统性能的提高并没有达到令人满意的程度，而且特征向量多数趋于离散化；自适应选择方法的改进以及如何降低预测时间，这些都是我们下一步研究的主要方向。

参 考 文 献

[1] Li W, Koendjbiharie W, de M Juca R C, et al. Algorithm for estimating bus arrival times using GPS data[C]// Proceedings of the 5th IEEE International Conference on Intelligent Transportation Systems. Singapore, 2002: 868-873

[2] 李天雷. 基于GPS数据的公交行程时间计算与预测系统[D]. 长春：吉林大学，2009

[3] Li F, Yu Y, Lin H, et al. Public bus arrival time prediction based on traffic information management system[C]// 2011 IEEE International Conference on Service Operations, Logistics, and Informatics (SOLI). 2011: 336-341

[4] Chung E H, Shalaby A. Expected time of arrival model for school bus transit using real-time global positioning system-based automatic vehicle location data [J]. Journal of Intelligent Transportation Systems: Technology, Planning, and Operations, 2007, 11(4): 157-167

[5] Liu H, Zuylen H V, Lint H V, et al. Predicting urban arterial travel time with state-space neural networks and Kalman filters [C]// Transportation Research Board, Annual Meeting (CD-ROM). Washington, 2006: 99-108

[6] Park J, Chen Z, Kiliaris L, et al. Intelligent vehicle power control based on machine learning of optimal control parameters and prediction of road type and traffic congestion[J]. IEEE Transactions on Vehicular Technology, 2009, 58(9): 4741-4756

[7] Yu B, Lam W H K, Tam M L. Bus arrival time prediction at bus stop with multiple routes [J]. Transportation Research Part C: Emerging Technologies, 2011, 19(6): 1157-1170

[8] Shen X, Chen J. Study on prediction of traffic congestion based on LVQ neural network [C] // International Conference on Measuring Technology and Mechatronics Automation. 2009: 318-321

[9] Rahman H A, Marti J R, Srivastava K D. Road traffic forecasting through simulation and live GPS-Feed from intervehicle networks[C]// Global Humanitarian Technology Conference (GHTC). 2012: 36-40

[10] 于滨，杨忠振，林剑艺. 应用支持向量机预测公交车运行时间[J]. 系统工程理论与实践，2007，27(4)：160-164

[11] 孙玉砚，刘燕，周新运，等. 基于路况相似性的城市公交车到站时间预测机制[J]. 软件学报，2012，23(zk1)：87-99

[12] 曹琛荔. 基于神经网络的智能公交信息服务系统研究[D]. 武汉：武汉理工大学，2011