# 基于概念格的概念相似度计算*)

智慧来[1] 智东杰[2] 刘宗田[1]

(上海大学计算机工程与科学学院 上海 200072)[1] (河南理工大学计算机科学与技术学院 焦作 454150)[2]

**摘 要** 为了提高信息检索的查全率和查准率,经常要处理相似的概念,因此计算概念间的相似度是必要的。概念由对象和属性两部分组成,所以基于概念格相似度计算的也分为两部分:一是利用概念格的层次关系计算对象的相似度,另一部分计算属性的相似度。其次,概念间的相似度与概念在概念格中的深度有关,计算时利用深度对结果加以修正。计算模型利用了概念较为完整的信息,计算结果与人工判断基本吻合。

**关键词** 概念格,概念相似度,对象,属性,深度

## Concept Similarity Based on Concept Lattice

ZHI Hui-lai[1] ZHI Dong-jie[2] LIU Zong-tian[1]

(School of Computer Engineering and Science, Shanghai University, Shanghai 200072, China)[1]

(School of Computer Science and Technology, He'nan Polytechnic University, Jiaozuo 454150, China)[2]

**Abstract** In order to improve the recall and precision, similar concepts are often processed during the information retrieval. With this regard, the possibility of evaluating concept similarity is acquiring an increasing relevance, since it allows the identification of different concepts that are semantically close. For every concept consist of two parts which include objects and attributes, so the computing model also has two parts. First part deals with the objects similarity by using hierarchy structure of concept lattice. The second part deals with the attributes similarity by using the expert's knowledge. Beside this, the depth also has relevance with concepts similarity. So the result is altered with concepts' depth. This model uses concept's integral information, and the result fits the judge given by expert.

**Keywords** Concept lattice, Concept similarity, Object, Attribute, Depth

领域本体和形式概念分析的目的都是在于概念的建模,虽然它们有各不相同的目的。与此同时,形式概念分析在本体工程中的应用越来越引起人们的重视,成为一个具有美好前景的研究领域。正由于这些原因,概念相似性的研究已越来越引起人们的关注。

近一两年内,国内外有一些学者对概念相似性的度量进行了较为深入的研究。文献[1]利用概念间的距离和最近根概念深度来计算概念相似度,并设计了基于语义相似度的信息检索方法。文献[2]中具体给出两种词汇语义相似度计算方法,其中第一种方法计算词汇语义相似度基于词语间距离度量,第二种方法计算词汇语义相似度,则是建立在义原相似度基础上。文献[3]利用形式概念分析来计算概念的相似度,但忽略了概念的深度对相似度的影响。

## 1 形式概念分析

R. Wille[4]提出的形式概念分析是以序理论和完备格理论为基础,依据数据库中提供的基本信息建立起的一种刻画对象与属性之间关系的数学结构。这种概念及概念层次的数学化,使形式概念分析成为数据挖掘与知识发现的重要方法,并广泛应用于许多领域[5-8]。

**定义 1** 设 $K=(U,A,I)$是一个形式背景,$X\subseteq U,B\subseteq A$。如果 $X,B$ 满足条件 $X'=B,B'=X$,则称序对$(X,B)$为形式背景 $K$ 的一个概念。$X$ 称为概念$(X,B)$的外延,$B$ 称为概念$(X,B)$的内涵[4]。$L(U,A,I)$或 $L(K)$表示 $K$ 中所有概念全体构成的集合,即

$$L(U,A,I)=\{(X\times B)\in U\times A;X'=B,B'=X\}$$

**定义 2** 设 $K=(U,A,I)$是一个形式背景,$(X_1,B_1)$,$(X_2,B_2)\in L(K)$,如果 $X_1\subseteq X_2$或$(B_2\subseteq B_1)$,称$(X_1,B_1)$是$(X_2,B_2)$的子概念,记为 $X_1,B_1)\leqslant(X_2,B_2)$。显然 $L(K)$关于"$\leqslant$"构成一个格,称为概念格[4]。

对于给定形式背景 $K=(U,A,I)$的两个概念$(X_1,B_1)$,$(X_2,B_2)$,以下结论成立:

If $X_1\subseteq X_2$ then $X'_2\subseteq X'_1$ for $X_1,X_2\subseteq U$; If $B_1\subseteq B_2$ then $B'_2\subseteq B'_1$ for $B_1,B_2\subseteq A$。

从概念格的定义和性质可以看出,概念格不仅精确定义了概念,更重要的是描述了概念间的继承关系,这为度量概念的相似性提供了良好的数据结构。

## 2 概念相似度计算模型

### 2.1 概念相似度计算的基本观点

概念相似度可以由概念在概念格层次结构中的距离来度量,直观上可以看出:距离越大,其相似度越低;反之,两个概念距离越小,其相似程度越大。距离为 0 时,其相似度为 1;概念距离为无穷大时,其相似度为 0;相似度为概念距离的单调递减函数。

概念所处层次的深度对相似度计算也有影响,概念相似

*)国家自然科学基金(项目批准号 60575035)。智慧来 博士生,主要研究领域为为信息处理、概念格、本体;智东杰 研究员,主要研究方向为人工智能、符号计算;刘宗田 硕士,教授,博士生导师,主要研究领域为人工智能、软件工程和形式概念分析。

度随着它们所处层次深度的增加而增加。因为，层次深度的增加意味着分类趋向细致，概念之间的相似程度就越高。

### 2.2 概念相似度计算

概念间的距离在概念格中可以用对象和属性的相似度来度量：两个概念距离越大，两个概念相同的对象和属性个数就越少；反之，两个概念距离越小，两个概念相同的对象和属性个数就越多。

在文献[3]中，计算属性相似度时，不仅计算了相同属性的相似度，不同属性的相似度也假定可能具有某种程度的相似。这种假定在一个统一的知识背景中是不恰当，属性就是用来区分不同的对象。如果属性相似，那么这个属性就不具有区分能力；因此，任何两个属性是不可能有交叉的语义的，在计算时考虑相同的属性即可。

**定义 3**(概念相似度 Sim)　在一个领域本体和若干个形式背景$(U_i,A_i,I_i)$　$i=1\cdots k$下，两个概念$(X_1,B_1),(X_2,B_2)$的相似度 sim 定义如下：

$$\mathrm{Sim}((X_1,B_1),(X_2,B_2))=\left(\frac{|X_1\cap X_2|}{m}*a+\frac{|B_1\cap B_2|}{n}*b\right)*(1+c)^{(l_1+l_2)}$$

其中，$a+b=1$，$c>0$，$m=\max(|X_1|,|X_2|)$，$n=\max(|B_1|,|B_2|)$

根据概念格的对偶原理[4]，概念的对象和属性具有同等的地位，可取$a=b=0.5$；$l_1,l_2$是$(X_1,B_1),(X_2,B_2)$在概念格中的层次；$c$是为了体现概念深度对相似度的影响而作的修正，在此取$c=0.01$。

### 2.3 计算实例

计算实例中的形式背景来源于文献[3]。这个形式背景描述了欧洲的一些著名城市：Athens(A)，Courmayeur(C)，Innsbruck(I)，London(L)，Paris(P)，Reykjavik(Re)，Rome(Ro)，它们具有属性：Archeological-site(Arc)，Beach(Bea)，Capital(Cap)，Euro(Eur)，River(Riv)，Skiing-area(Ski)。由形式背景建立的概念格如图 1。

表 1　欧洲城市的形式背景

| | Arc | Bea | Cap | Eur | Riv | Ski |
|---|---|---|---|---|---|---|
| A | * | * | * | * | | |
| C | | | | * | | * |
| I | | | | * | * | * |
| L | | | * | | * | |
| P | | | * | * | * | |
| Re | | | * | | | * |
| Ro | * | * | * | * | * | |

根据定义 3，可以计算多对概念的相似度如下：

兄弟概念的相似度 Sim(({L,P,Ro},{Cap,Riv}),({A,P,Ro},{Cap,Eur}))=0.61

父子概念的相似度 Sim(({L,P,Ro},{Cap,Riv}),({P,Ro},{Cap,Eur,Riv}))=0.70

祖孙概念的相似度 Sim(({P,Ro},{Cap,Eur,Riv}),({A,L,P,Re,Ro},{Cap}))=0.38

没有直接亲缘关系的概念相似度 Sim(({Ro},{Arc,Bea,Cap,Eur,Riv,Ski}),({I},{Eur,Riv,Ski}))=0.21

从计算结果可以看出，兄弟概念的相似度小于父子概念的相似度，而祖孙概念的相似度更小。祖孙概念的相似度大于没有直接亲缘关系的概念间的相似度。

多个形式背景可以通过合并形成单个形式背景[4]，生成概念格后再进行概念相似度的分析，这里不再赘述。

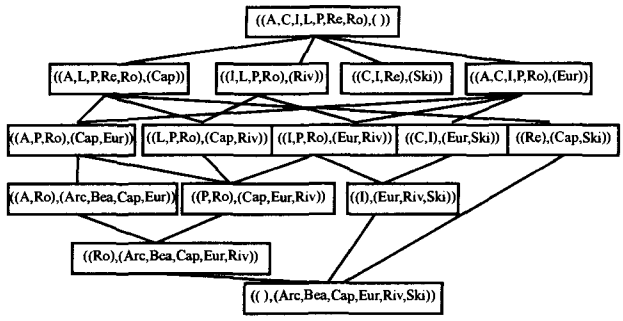


图 1　“欧洲城市”形式背景的概念格

## 3　应用与结论

在信息检索的实现过程中为了高查全率，如果按照查询关键字进行检索得不到相关信息，那么就对查询关键字进行泛化和平级扩展操作[9]，用其相似概念进行查询。

检索算法描述如下：

```
Result Retrieve(Key)
{
   Find relative data and save to Result;
     If Result! =Null
   Then return Result;
   else
{ New_Key=Find_Parent_Concept(Key);
   If Sim(New_Key,Key)>ε//计算相似度
     Then
{Key=New_Key;
 Retrieve(Key);} //递归调用
};//泛化操作
{New_Key=Find_Sibling_Concept(Key);
  If Sim(New_Key,Key)>ε //计算相似度
    Then
{Key=New_Key;
 Retrieve(Key);} //递归调用
};//平级扩展操作
}
```

结合笔者的研究领域，建立了一个关于“本体”的领域本体，包括：本体、事件、主题、基于事件的本体、领域本体等多个概念形成的形式背景。结合上述算法并结合设计了一个文献检索的工具，并在从“中国知网”下载的关于本体的 2463 篇文献中进行实验。实验结果表明提高了检索的查全率。例如：没有应用算法 Retrieve 以前，输入“事件本体”进行文献检索，没有文献命中，应用算法之后返回了 14 篇文献，其中有 3 篇与我们所要研究的“事件本体”不相关。

实验表明，本文中的计算模型充分考虑到了距离对概念相似度的影响，并利用概念格特有的层次结构，计算模型简单明了，计算结果能够满足实际应用的要求。

## 参 考 文 献

[1] 王进，陈恩红，施德明，等. 一种基于语义相似度的信息检索方法. 模式识别与人工智能[J]，2006，12(6)：696-701

[2] 吴健，吴朝晖，李莹，等. 基于本体论和词汇语义相似度的 Web 服务发现. 计算机学报[J]，2005，4(4)：595-602

(下转第 167 页)

②$D_1|_X \neq D_2|_X$ 且 $D_1|_Y \doteq D_2|_Y$，又由 $D \models Y \rightarrow Z$，则 $D_1|_Y \doteq D_2|_Y$ 时，$D_1|_Z \doteq D_2|_Z$ 成立，故 $D \models X \rightarrow Z$ 成立。

由上述讨论知，传递规则成立。

(4) 设 $\forall D_1, D_2 \in \{D|_{XYWZ}\}$，由 $D \models X \rightarrow Y$ 得 $D_1|_X \doteq D_2|_X$ 时，$D_1|_Y \doteq D_2|_Y$ 成立，又由 $D \models WY \rightarrow Z$ 得 $D_1|_{WY} \doteq D_2|_{WY}$，$D_1|_Z \doteq D_2|_Z$ 成立，对 $D_1|_{WX}$ 和 $D_2|_{WX}$ 分以下两种情况讨论：

1) $D_1|_X \neq D_2|_X$，则 $D_1|_{WX} \neq D_2|_{WX}$，此时无论 $D_1|_Z$ 和 $D_2|_Z$ 取何值时，$D \models WX \rightarrow Z$ 成立；

2) $D_1|_X \doteq D_2|_X$，由 $D \models X \rightarrow Y$ 得 $D_1|_X \doteq D_2|_X$ 时，$D_1|_Y \doteq D_2|_Y$ 成立。又有两种可能：

①若 $D_1|_W \doteq D_2|_W$，则 $D_1|_{WY} \doteq D_2|_{WY}$，又由于 $D \models WY \rightarrow Z$ 得，所以当 $D_1|_{WY} \doteq D_2|_{WY}$ 成立时，$D_1|_Z \doteq D_2|_Z$ 也成立。由于 $D_1|_X \doteq D_2|_X$、$D_1|_W \doteq D_2|_W$，则有 $D_1|_{WX} \doteq D_2|_{WX}$。于是有 $D_1|_{WX} \doteq D_2|_{WX}$ 时，$D_1|_Z \doteq D_2|_Z$ 成立，所以 $D \models WX \rightarrow Z$ 成立；

②若 $D_1|_W \neq D_2|_W$，则 $D_1|_{WX} \neq D_2|_{WX}$，此时无论 $D_1|_Z$ 和 $D_2|_Z$ 取何值时，$D \models WX \rightarrow Z$ 都成立。

由上述讨论知，伪传递律成立。证毕。

**定义 12** 设 $D \models S$，$F$ 表示通过 $S$ 的 $XSFD$ 集，若 $D \models F$ 成立，$D \models X \rightarrow Y$ 也成立，则称 $F$ 逻辑蕴涵 $X \rightarrow Y$，记作 $F \models X \rightarrow Y$。

**定义 13** 关于 $R$ 通过 $S$ 的 $XSFD$ 集 $F$ 的闭包是由 $F$ 根据 $R$ 推出的所有 $XSFD$ 的集合，记作 $F^+ = \{X \rightarrow Y | F|\text{-}X \rightarrow Y\}$；$X \subseteq S$，子树 $X$ 的闭包记作 $X_F{}^+ = \cup\{Y |$ 存在 $X \rightarrow Y \in F^+\}$。

**定理 3** 由推理规则 R 构成的公理系统是完备的。

证明：设 $F$ 表示通过 $S$ 的 $XSFD$ 的集合，$X \rightarrow Y \in F$。根据子树规则，$Y = \{A\}$，$A$ 表示单一路径。

(反证法)。假设 $X \rightarrow A \in F^+$，但是不能通过 $F$ 使用推理规则而推出。为了证明推理规则是完备的，需要构造出一棵 $D \models S$，使 $F$ 在 $D$ 上成立，同时使 $X \rightarrow A$ 在 $D$ 上不成立，这样就与 $X \rightarrow A \in F^+$ 矛盾。

(1)构造 $D$。设 $D \models S$，$\{0,1\}$ 表示 $D$ 中叶子节点的两个不同的非空值，$X(S, \forall D_1, D_2 \in \{D|_S\}$，满足 $\forall W \in X$，$val(D_1|_W) = \perp$ 且 $\forall W \in S\text{-}X$，$val(D_1|_W) = 1$；$\forall W \in S$，$val(D_2|_W) = 0$。$D_1$ 和 $D_2$ 分别如图 1 和图 2 所示，$r$ 表示根节点，$\perp$ 的取值范围为 0。

(2)修改 $D$，判定 $F$ 是否成立。对于任意的 $XSFD$ $W \rightarrow B \in F$，则修改 $val(D_1|_B) = val(D_2|_B) = 0$，由定义 11，此 $XSFD$ $W \rightarrow B$ 在修改后的文档树 $D$ 中成立。

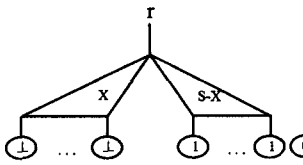


图 1 文档树 $D_1$

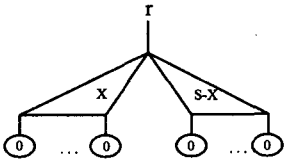


图 2 文档树 $D_2$

(3)判定 $XSFD$ $X \rightarrow A$ 在 $D$ 上成立的条件。

当 $val(D_1|_A) = 0$ 时，$val(D_1|_A) = val(D_2|_A)$，即 $D_1|_A \equiv D_2|_A$，又由 $D_1|_X \doteq D_2|_X$，则 $X \rightarrow A$ 成立。若 $X \rightarrow A$ 成立，$D_1|_X \doteq D_2|_X$ 时，则 $D_1|_A \doteq D_2|_A$，所以 $val(D_1|_A) = 0$。

所以 $val(D_1|_A) = 0$ 是 $X \rightarrow A$ 成立的充要条件。

(4)根据(2)，任意的 $W \rightarrow Z \in F$，则 $D \models W \rightarrow Z$。又由于 $F \not\models X \rightarrow A$，则 $A \notin X^+$，再由(3)，$val(D_1|_A) \neq 0$，此时 $X \rightarrow A$ 在 $D$ 上不成立。综合可得推理规则集是完备的。证毕。

**结束语** 在 XML 文档中具有不完全信息的情况下，本文研究了 XML 强函数依赖理论。通过对 XML 强函数依赖理论的研究，对 XML 文档中的不确定性数据进行了约束，使维持 XML 数据库的完整性操作更加容易。

## 参 考 文 献


[1] 郝忠孝. 空值环境下数据库导论. 机械工业出版社，1996

[2] Arenas M, Libkin L. A normal form for XML documents. ACM Transactionson Database Systems, 2004, 29(1): 195-232

[3] Wang Junhu, Topor R. Rkemoving XML data redundancies using functional and equality generating dependencies. In Database Technologies ADC. Newcastle, Australia, 2005

[4] Vincent M W, Liu Jixue. Strong functional dependencies and a redundancy free normal form for XML. In Systemics, Cybern-etics and Informatics-SCI, 2003, IIIS: 218-22

[5] Vincent M W, Liu Jixue. Strong functional dependencies and their application to normal forms in XML. ACM Transactions on Database Systems, 2004, 29(3): 455-462

[6] Levene M, Loizu G. Axiomatisation of functional dependencies in incomplete relations. Theoretical computer science, 1998, 206(1/2): 283-300

[7] Hartmann S, Link S, Kirchberg M. A subgraph-based approach to-wards functional dependencies for XML. In Systemics, Cybern-etics and Informatics-SCI, 2003, IIIS: 200-205


(上接第 157 页)


[3] Formica A. Ontology-based concept similarity in Formal Concept Analysis. Information Science[J], 2006, 176: 2624-2641

[4] Ganter B, Wille R. Formal Concept Analysis: Mathematical Foundation[M]. New York. Springer-Verlag, 1999

[5] Dtintsch I, Gediga G. Algebraic aspects of attribute dependencies in information systems. Fundamenta Informaticae[J], 1997, 29: 119-133

[6] Dtintsch I, Gediga G. Approximation operators in qualitative data analysis // de Swart H, Orlowska E, Schmidt G, et al., eds. Theory and Application of Relational Structures as Knowledge Instruments[M]. Springer, Heidelberg, 2003: 216 - 233

[7] Pagliani P. From concept lattices to approximation spaces: Algebraic structures of some spaces of partial objects. Fundamenta In Formaticae[J], 1993, 18(1): 1-25

[8] Yao Y Y. Concept lattices in rough set theory //Proceedings of 23rd International M eeting of the North American Fuzzy Information Processing Society. 2004

[9] 曹锐，陈刚，蔡铭. 基于本体的网络化制造资源检索. 计算机工程[J], 2004, 2(3): 143-146