基于粗糙集的分类关联规则挖掘算法研究

尹世群　余建桥　葛继科　邱玉辉

（西南大学计算机与信息科学学院　重庆 400715）

摘　要　本文给出了一种将属性约简和分类关联规则挖掘相结合的新型分类挖掘系统的算法(CARMA)。它运用粗糙集理论把关系数据库按属性值分成若干等价类、约简冗余属性及依赖属性，然后对数据约简后的目标关系表求取分类支持度大于阈值的强类和特征置信度大于阈值的强特征，从而有效获取强类中的强特征的决策关联规则。实验结果表明，CARMA 对于数据的分类是有效的，比其它算法具有更高的分类精度和效率。它能够有效地克服 ID3 系列算法的冗余性、复杂性和对大数据量的不适应性，对增量数据能够达到较好的分类效果和具有广泛的应用前景。本文关键讨论了具体的算法、系统框架和实例。

关键词　粗糙集，属性约简，分类挖掘，分类支持度，特征置信度

Study of Classification Association Rule Mining Algorithm on Rough Set

YIN Shi-Qun　YU Jian-Qiao　GE Ji-Ke　QIU Yu-Hui

(Faculty of Computer and Information Science, Southwest University, Chongqing 400715)

Abstract　This paper brings up a new classification data mining system algorithm (CARMA) combined attribute reduction and classification mining of association rule. Based on Rough set theory it divides relation table into several equivalence class according to attribute values, reduces redundancy attribute and dependence attribute, and then gets strong-class whose support degree of categories are more than threshold values and strong-characteristic whose confidence degree are more than threshold values from the relation table being reduced data. At last we can get decision-making association rules of strong-characteristic in the strong-class. Its experimentation result makes know that the CARMA data classification is valid and has higher classification accuracy and efficiency than other method. In this way the redundancy and the complexity of ID3 series algorithm can be reduced effectively, and it gets better classification effect to increasing data. It has widespread application perspective of the most large or increment relation databases mining. The algorithm, system framework and example are discussed in details.

Keywords　Rough set, Attribute reduction, Classification mining, Support degree of category, Confidence degree of characteristic

1　引言

分类即区分数据类别，是数据挖掘中应用最多的任务。首先从数据中选出已经分好类的训练集，在此训练集上运用分类技术，建立用规则或决策树表示的分类模型，即找出一个类别的概念描述。然后，根据分类模型对于没有分类的数据进行分类。建立分类决策树的方法，典型的有 ID3、C4.5、IBLE 等方法。建立分类规则的方法，典型的有 AQ 方法、粗糙集方法、遗传分类器等。

其中建立分类决策树的方法以 ID3(Iterative Dichotomizer 3)系列算法(包括 ID3、C4.5、C5.0)为代表，就是将分类结果以分类树的形式给出，树的内部节点是一个决策，而叶节点代表一个类，而不是实际的规则[1]。进行分类时要查找树，而且算法产生时比较复杂，要求将分类树放入内存，进行查找。在产生的分类树转换成规则过程中，考虑了所有的因素，对主要的因素和次要的因素都要作出选择，不利于规则的生成和化简。以属性为分类节点的 ID3 为代表的一类算法的效率对于较少的数据而言是适当的，但是随着数据量的增加和决策属性的增加，则效率会大幅度下降，而且不容易形成规则[1]。

而粗糙集(Rough sets)理论[2]是波兰数学家 Z. Pawlak 在 1982 年首先提出的一种用于处理不确定性和含糊性知识的数学工具，目前在数据挖掘的各方面已有很好的应用，其基本思想是在保持分类能力不变的前提下，通过知识约简，导出概念的分类规则。它无需提供相关数据集合外的任何先验信息，适合于发现数据中隐含的、潜在有用的规律，即知识，找出其内部数据的关联关系和特征[3]。

其中属性约简是 Rough 集理论的核心内容之一，也通常作为数据挖掘的一个预处理步骤。属性约简是通过求属性重要性并排序，在泛化关系中可以找到一个较小的属性集 $B \subseteq A$，使得可用 A 描述的对象集合必然可用 B 描述，从而去除了不必要的属性，实现信息约简，简化了分类的标准[4]。约简包括属性约简和属性值约简。目前，研究的重点放在属性约简方面，提出了多种属性约简算法，如基于正区域的属性约简算法[5]，基于区分矩阵的属性约简算法[6]，基于信息熵的属性约简算法[7]等等。属性值约简的研究也是基于这些算法。目前这些算法将重点放在约简的完备性。约简作为一种新的数据挖掘方法，已应用在"海量"数据表中。通常情况下，我们不关心那些出现几率很小的决策规则。从上个世纪 80 年代以来，在统计模式识别、机器学习和数据挖掘[3]等领域，属性约简已经成为一个研究与开发成果丰富的领域。

尹世群　副教授，博士生，主要研究方向：人工智能、数据挖掘、Web 信息处理等；余建桥　教授，博士，主要研究方向：人工智能、智能信息处理等；葛继科　博士生，主要研究方向：人工智能、智能信息处理等；邱玉辉　教授，博士生导师，主要研究方向：自动推理、机器学习、人工智能等。

如果用粗糙集方法建立分类模型，会大大地方便用户并能提高数据的分类精度和效率。为此，本文将属性约简和分类关联规则[8~10]挖掘相结合，给出了分类支持度、特征置信度、属性重要性等概念，在此基础上，提出了一种新的基于粗糙集的分类挖掘系统的算法(CARMA)。它关键运用粗糙集理论把关系数据库按属性值分成若干等价类、约简冗余属性及依赖属性，然后对数据约简后的目标关系表求取分类支持度大于阈值的强类和特征置信度大于阈值的强特征，从而获取强类中的强特征的关联知识规则。它能够有效地克服ID3系列算法的冗余性和复杂性，提高数据的分类精度和效率。这一分类技术的实现方法是以数据库中关系表为基础，而且在原始数据增加的情况下，可以通过约简来压缩数据规模，使之只与属性值有关系，而与原始的数据量无关。而现在的数据存放中，几乎所有的数据都是用关系表存放，这为发现属性间的联系形成决策规则、构造挖掘系统提供了极大方便。

2 分类挖掘系统框架

分类挖掘系统框架如图1所示。其中，第一部分为属性约简，在各个分类的训练数据库上使用属性约简去除冗余属性和依赖属性，确定最小属性集；第二部分为分类关联规则挖掘，用挖掘算法来挖掘，求取分类支持度大于阈值的强类和特征置信度大于阈值的强特征，从而获取强类中的强特征的决策规则，形成一个分类规则库；第三部分为分类引擎，对于一个输入待分类数据，分类器利用一个规则匹配算法对该数据进行匹配和分类，并输出分类结果。

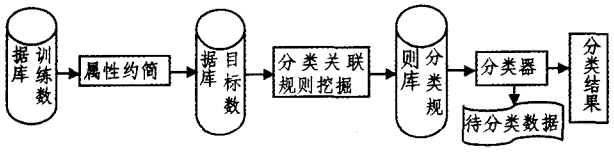

图1 分类挖掘系统框架的流程图

3 基于粗糙集的属性约简

3.1 粗糙集的基本概念

粗糙集理论中的基本概念是其处理思想和算法的基础，下面介绍几个主要概念：信息系统、不可分辨关系和等价关系与划分、近似集、正区域及约简、属性的依赖度及属性的重要性等。

定义1 一个信息系统 S 可以表示为 $S=\langle U, A, V, f\rangle$，其中 U 是对象的非空有限集合，即论域；A 是非空有限属性集合；$V=\bigcup_{a\in A} v_a$，V_a 表示属性 a 的值域；$f: U\times A\to V$ 是一个信息函数，它指定 U 中每一个对象 x 的属性值，即对 $x\in U, a\in A$ 有 $f(x,a)\in V$。如果属性集 A 可以分为条件属性集 C 和决策属性集 D，即 $C\cup D=A, C\cap D=\phi$，则该信息系统称为决策系统或决策表，决策系统是一类最为常见的信息系统。

定义2 在信息系统 $S=\langle U, A, V, f\rangle$ 中，对于每个属性子集 $B\subseteq A$，可以定义一个不可分辨关系 IND(B)：

$$\mathrm{IND}(B)=\{(x,y)\in U\times U \mid b\in B, f(x,b)=f(y,b)\}$$

不可分辨关系是一种等价关系，它把 U 划分为有限个集合，称为等价类，在每个集合中，对象间是不可分辨的。针对属性集 B 上的不可分辨关系，U 可划分为几个等价类，用 $U/IND(B)$ 表示。

对象 $x\in U$ 在属性集 B 上的等价类：

$$[x]_B=\{y \mid (x,y)\in \mathrm{IND}(B)\}$$

容易看出，在信息系统中，一个属性对应一个等价关系，一个表可以看作是定义的一族等价关系，即知识库。

定义3 给定一个信息系统 $S=\langle U, A, V, f\rangle$，对于任意一个对象集合 $X\subseteq U$ 以及属性集合 $B\subseteq A$。

X 的 B 下近似定义为：$\underline{B}X=\{x \mid [x]_B\subseteq X\}$；

X 的 B 上近似定义为：$\overline{B}X=\{x \mid [x]_B\cap X\neq\phi\}$

定义4 对于决策系统 $S=\{U, CUD\}$，C 为条件属性集合，D 为决策属性集合。$B\subseteq C$，定义 B 相对于 D 的正区域为

$$POS_B(D)=\{\underline{B}X \mid X\in U/IND(D)\}$$

$POS_B(D)$ 实际上是那些可以根据属性集合 B 准确地被分入由属性 D 所确定的分类元素的集合。

设 $a\in C$，$POS_C(D)=POS_{C-\{a\}}(D)$，则称 a 为 C 中可省略。当 C 中每个元素都不为 C 中 D 可省略时，称 C 为 D 独立。当 $C'=C-C*$ 为 D 独立，且 $C*$ 中的所有元素都是 D 可以省略的话，则称 C' 为 C 的 D 相对约简。从分类的角度来理解，相对约简就是一种分类来表达另一种分类必不可少的属性集合[11]。

定义5 在属性约简中，利用二个属性集合 $R\subseteq C, P\subseteq D$ 之间的相互依赖程度，可以定义一个属性 a 的重要性。属性集 P 对 R 的依赖度用 $\gamma_R(P)$ 表示。其定义如下：

$$\gamma_R(P)=\frac{card(POS_R(P))}{card(U)}$$

其中 $card(.)$ 表示集合的基数。

不同属性对于决定条件属性和决策属性之间的依赖关系起着不同的作用。对于 $a\in R$，对分类 $U/IND(P)$ 重要程度定义为：$SGF(a,R,P)=\gamma_R(P)-\gamma_{R-\{a\}}(P)$[12]。

定理1 如果在条件属性集 C 中存在在依赖关系 $\{C_i, C_j, C_k\}\to\{C_m\}$，且 $\gamma_{\{G,G_j,\alpha\}}(C_m)=1$，那么在条件属性集 C 中删除属性 C_m，即 $C'=C-\{C_m\}$，决策属性与条件属性集的依赖程度不变，即有 $\gamma_C(D)=\gamma_{C'}(D)$。

根据定理，如果 $\gamma_{\{G,G_j,\alpha\}}(C_m)=1$，则 C_m 为可约简属性。

确定最小条件属性集的方法就是从条件属性中删除可约简属性、有冗余属性(与决策属性无关同时又不依赖其它条件属性的条件属性)和依赖属性(依赖于其它条件属性的条件属性)。

定义6 条件属性集中可删除属性的候选集 cand_set 定义为

$$\mathrm{Cand\_set}=\{C_i\in C \mid \gamma_C(D)=\gamma_{C-\{G\}}(D)\}$$

定理2 如果 $C_i\in$ cand_set，且 $\gamma_{C-\{C_i\}}(C_i)\neq 1$，则 C_i 为冗余属性。

有些属性依赖于冗余属性，当冗余属性从属性集中删除后，这些属性不再依赖于其它属性，因此并非所有 cand_set 中的属性都是可删除的。

3.2 属性约简算法

目前属性约简的算法很多，大多数是利用分辨矩阵来计算核心，作为计算约简集的起点，逐步找到最小属性约简的集合。由于用到二维矩阵，这种算法时间和空间复杂度一般较高。本文提出的属性约简算法为了避免使用分辨矩阵，从空集出发，利用属性重要度作为启发式信息找出可删除属性候选集，然后在候选集中确定冗余属性，并在属性集中将之删除，在新的属性集中再确定依赖属性并在属性集中将之删除，这样可快速地求得一个最小条件属性集。基于粗糙集的属性约简算法如下。

算法 1

输入:经过数据预处理及离散化后的决策表 S=〈U, CUD, V, f〉,其中 C 为条件属性集,D 为决策属性集。
输出:决策表 S=〈U, C'UD, V, f〉,C'为与 D 存在依赖关系的最小条件属性集。

```
Begin
(1) //计算可删除属性候选集 cand_set
    Cand_set=Φ;
    For all Ci∈C do
        If γC(D)=γC-{G}(D) then
          Cand_set= Cand _set∪{Ci};
        Eedif
    Next
(2) //如果 Cand_set≠Φ,删除冗余属性
    If Cand_set≠Φ
      C'=C
      For all Ci∈cand_set do
          If γC'-{Ci}(C')≠1 then
            cand_set= cand_set-{Ci};
            C'= C'-{Ci};
          Endif
      Next
    Endif
(3) //如果 Cand_set≠Φ,删除依赖属性
    If Cand_set≠Φ
          For all Ci∈cand_set do
            If γC-{G}(D)=γC(D) then
              cand_set= cand_set-{Ci};
              C'= C'-{Ci};
            Endif
          Next
      Endif
    End.
```

4 分类关联规则挖掘算法

分类规则挖掘是对具有最小条件属性集约简表进行知识发现,找出元组所具有的特征,然后以规则式来描述挖掘出的知识。

定义 7 设 S 是一个 4 元组$\langle U,A,V,f\rangle$信息系统的决策表,S_r 是总记录数,B 是 A 的子集,X 是关于 B 的一个等价类,S_x 是 X 的记录个数,则称 $S=S_x/S_r*100\%$ 是等价类 X 的分类支持度。

对于关系表 S,按照某个属性集 B 进行分类分成等价类 $B_1,B_2,\cdots$,表示为 $IND(B)=\{B_1,B_2,\cdots\}$,可以求出每个等价类对应的分类支持度$\{S_1,S_2,\cdots\}$。

定义 8 设 S_t 是一个给定的阈值,$0\leqslant S_t\leqslant 1$,对于分类支持度 $S\geqslant S_t$ 的等价类,称为强类,反之则称为弱类。

在大量的数据中,常常关心和感兴趣的是分类支持度较大的强类。强类所包含的知识更具有代表性,需对强类作进一步的特征分析,然后发现其隐含的规则。

定义 9 设 E 是基于条件属性 C 的一个等价类,H 是决策属性 D 的子集,基于 H 的等价类 F 称为 E 分类中的特征域,H 中各属性取值$\{h_1,h_2,\cdots\}$称为 E 分类中的特征。

定义 10 设 Ec 是等价类 E 中的记录数,Fc 是特征域 F 中的记录数,则称 $C=F_c/E_c*100\%$为特征置信度。

定义 11 设 C_t 是一个给定的阈值,$0\leqslant C_t\leqslant 1$,特征置信度 $C\geqslant C_t$ 的特征域称为强特征域,反之则称为弱特征域,强特征域中属性取值称为强特征,弱特征域中属性的取值为弱特征。

强类 E 中的强特征 f 是具有代表性的知识,表达为关联规则 $E\rightarrow f$。

获取分类关联规则的关键是求强类及其上的强特征。其具体实现算法为

算法 2

输入:决策表 S=〈U, C'UD, V, f〉,C'为与 D 存在依赖关系的最小条件属性集。
输出:分类规则集。

```
Begin
(1) Ind(C')={E1,E2,…,En}
    //在约简后的决策表中,按条件属性 C'进行分类
(2) For i=1 to n Do
    //对每个等价类 Ei,Ei 若是强类就求取其强特征
      Si=Ei 的分类支持度;
      If Si≥St Then // St:分类支持度阈值
        求所有 Ei 分类中的特征域 F1,F2,…,Fk;
        Cj=Fj 的特征置信度  其中 1≤j≤k;
(3)     For j=1 to k Do
          If Cj≥ Ct Then //Ct:特征置信度阈值
            Fj 为强特征域;
            f=取 Fj 的强特征;
(4)         规则 Ei→f 放入结果库;
          Endif
        Next j
      Endif
    Next i
  End.
```

5 分类器的构建

用上述分类规则算法提取出所有有效的规则并形成一个规则知识库 $R=\{r_1,r_2,\cdots,r_q\}$,之后就可以构建分类器。该分类器包括数据库、规则库和分类模块。由于待分类数据可能和规则库中的许多序列相匹配,因此设计了一种有效的规则匹配度量机制,如图 2 所示。

通常从选择规则到执行规则分成三步:

第一步,把动态数据库的事实和知识库中的规则的条件部分相匹配;

第二步,当有一个以上的规则条件部分和当前事实相匹配时,进行冲突解决;

第三步,执行操作,即执行规则的操作部分。操作之后,动态数据库被修改,然后其它的规则又有可能被使用。

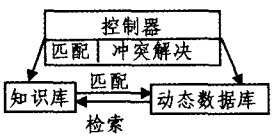

图 2 分类器的基本结构图

6 实验结果

为了验证 CARMA 的可行性和有效性,现用 CARMA 对天气进行分类。

下面用表 1 所示的经典天气决策表为例[13],来说明本文提出的算法的有效性。表 1 中,$a1$,$a2$,$a3$,$a4$ 是条件属性,分别代表天气、温度、湿度、风,d 是决策属性,论域 $U=\{x1,x2,\cdots,x14\}$。给定分类支持度阈值 $S_t=8\%$,特征置信度阈值 $C_t=50\%$。

表 1 关于天气的决策表

U	a1	a2	a3	a4	d
1	晴	热	高	否	N
2	晴	热	高	真	N
3	多云	热	高	真	N
4	雨	温暖	高	否	P
5	雨	冷	正常	否	P
6	雨	冷	正常	真	N
7	多云	冷	正常	真	P
8	晴	温暖	高	否	N
9	晴	冷	正常	否	P
10	雨	温暖	正常	否	P
11	晴	温暖	正常	真	P
12	多云	温暖	高	真	P
13	多云	热	正常	否	P
14	雨	温暖	高	真	N

利用上节给定的算法1进行约简和算法2来获取强特征，得到如下规则：

规则1 （a1，晴）∧（a3，高）→（d，N）。

规则2 （a1，多云）→（d，P）。

规则3 （a1，雨）∧（a4，否）→（d，P）。

规则4 （a1，雨）∧（a4，真）→（d，N）。

规则5 （a1，晴）∧（a3，正常）→（d，P）。

对于本文给出的CARMA与C4.5方法得到的性能对比如图3所示。可见，用CARMA进行分类的效率高。硬件平台为P4 2.8G Hz. 256M. 在SQL SERVER2000上实现。

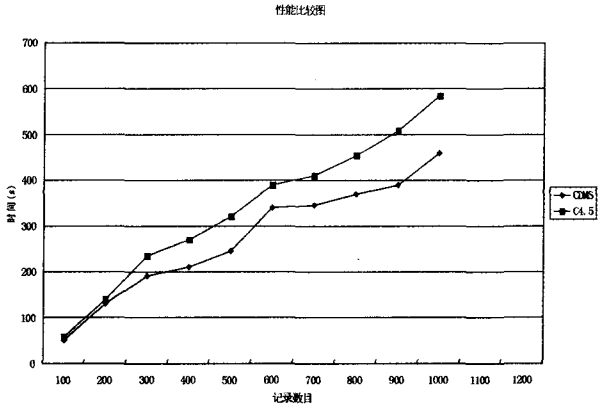

图3 CARMA与C4.5性能比较图

本文的属性约简算法1的空间复杂度是$O(|C|)$，而传统的算法用分辨矩阵计算约减属性的核，空间复杂度是$O(|C|^2)$，由此可见此算法大大降低了空间复杂度。

分类精确度定义为能够正确分类的记录数目和测试集中记录总数目的比值。实验对同样的数据集使用基于正区域的属性约简算法[5]、基于区分矩阵的属性约简算法[6]、基于信息熵的属性约简算法[7]和基于本文的属性约简算法，它们的分类精度实验结果分别是83.4、77.5、84.2、88.6。可见，用本文中的属性约简后分类的精确度比用其他几种属性约简方法更高。

结论 为了有效地、更快地对海量数据进行分类，本文提出了将属性约简和分类关联规则挖掘相结合的分类挖掘系统的算法。它运用粗糙集理论把关系数据库按属性值分成若干等价类、约简冗余属性及依赖属性，然后对数据约简后的目标关系表求取分类支持度大于阈值的强类和特征置信度大于阈值的强特征，从而有效获取强类中的强特征的决策规则。实验结果表明，CARMA对于数据的分类是有效的，目前比文中提到的其它方法具有更高的分类精度和效率。它能够有效地克服ID3系列算法的冗余性、复杂性和对大数据量的不适应性，而且在原始数据增加的情况下，可以通过约简来压缩数据规模，使之只与属性值有关系，而与原始的数据量无关，因此对增量数据能够达到较好的分类效果。算法的执行和处理是简单易行的，产生的规则也是准确的。领域决策者就可以直接运行分类器进一步对同类问题进行分类和决策。在各种评估工作、医疗诊断、交通信息、案件信息、天气预测等大型数据库的数据挖掘中有广泛的应用前景及实用价值。

参考文献

1 Han Jiawei, Kamber M. Data Mining: Concepts and Techniques. Simon Fraser University, 2000

2 Pawlak Z. Rough sets [J]. International Journal of Computer and Information Science, 1982, 11: 341～356

3 Mitra P, Murthy C A, Pal S K. Unsupervised feature selection using feature similarity[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(3): 301～312

4 张修平，仇国芳. 基于粗糙集的不确定决策[M]. 北京：清华大学出版社，2005. 28～37

5 Guan J W, Bell D A. Rough computational method for information systems [J]. Artificial Intelligence, 1998, 105(1-2): 77～103

6 Skowron A, Rauszer C. The discenibility matrices and fuction in information system [C]. In: Slowinski R, ed. Intelligent Decision Support Handbook of Application and Advance of The Rough Sets Theory, Dordrecht: Kluwer Academic Publishers, 1992. 331～362

7 苗夺谦，胡桂荣. 知识约简的一种启发式算法[J]. 计算机应用与发展，1999，36：681～684

8 Agrawal R, Imielinski T, A Swami. Mining association rules between sets of items in large database [C]. In: Proceeding of the ACM SOGMOD Conference on Management of data, 1993, 5: 207～216

9 Agrawal R, Srikant R. Fast algorithms for mining association rules [C]. In: Proceedings of the 20th VlDB Conference, Santiago, Chile, 1994. 487～499

10 Relue R, Wu Xindong, Huang Hao. Efficient Runtime Generation of Association Rules. In: Proceeding of 2001 ACM CIKM Tenth International on Information Knowledge Management, 2001

11 徐余法. 粗糙集理论及应用. 上海电机学院学报，2005，8(2)：39～43

12 蒋良孝，蔡之华，刘钊. 一种基于粗糙集的决策规则挖掘算法. 微机与应用，2004，3：7～9

13 瞿彬彬，卢炎生. 基于粗糙集的属性约简算法研究. 华中科技大学学报(自然科学版)，2005，33(8)：30～33

14 洪家荣. 归纳学习——算法理论和应用. 北京. 科学出版社，1997

15 潘巍，王阳生，杨宏戟. 粗糙集理论中求取最小决策规则的研究. 计算机科学，2007(5)

16 宋笑雪，解争龙，张文修. 集值决策信息系统的知识约简与规则提取. 计算机科学，2007(4)

17 孙成敏，刘大有，孙舒杨. 含序信息的粗集方法研究. 计算机科学，2006(11)

(上接第156页)

引入关系规则推导挖掘隐含的位置关系，并可利用其它关系知识交互验证位置关系。

参考文献

1 曹存根，张春霞，王海涛. 基于本体的文本知识获取研究[A]. 见：王珏，陆汝钤，等. 智能信息处理系列研讨会[C]. 上海：2003. 7～8

2 余蕾. 从大规模中文语料中获取和验证概念的研究[D]. 北京：中国科学院计算技术研究所，2006

3 于海滨，秦兵，刘挺，等. 命名实体识别和指代消解在文摘系统中的应用. 计算机应用研究，2006，23(4)：180～182，195

4 Guo Honglei, Jiang Jianmin, Hu Gang, et al. Chinese Named Entity Recognition Based on Multilevel Linguistic Features. In: IJCNLP-04, Hailan, China, March 2004. 294～231

5 Hearst M A. Automatic Acquisition of Hyponyms from Large Text Corpora. In: 14th International Conference on Computational Linguistics (COLING 1992), August 1992. 539～545

6 Pantel P, Ravichandran D, Hovy E. Towards Terascale Knowledge Acquisition. In: Proceedings of Conference on Computational Linguistics. (COLING-04), Geneva, Switzerland, 2004. 771～777

7 Cederberg S, Widdows D. Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction. In: Conference on Natural Language Learning (CoNLL-2003), Edmonton, Canada, 2003. 111～118

8 Hull R, Gomez F. Automatic acquisition of biographic knowledge from encyclopedic texts [J]. Expert Systems with Applications, 1999, 16: 261～270

9 宋柔，许勇. 基于词汇语义的百科辞典知识提取实验[J]. Computational Linguistics and Chinese Language Processing, 2002, 7(2): 37～40

10 宋柔. 汉语词汇语义信息的研究和应用[A]. 见：第三届中文词汇语意学研讨会论文集[C]. 台北：2002. 169～177

11 张春霞. 领域文本知识获取方法研究及其在考古领域中的应用[D]. 北京：中国科学院计算技术研究所，2005

12 毛汉英，刘伉. 世界人文地理手册. 北京：知识出版社，1983

13 顾芳. 多学科领域本体的设计方法研究[D]. 北京：中国科学院计算技术研究所，2004

14 刘磊，曹存根，王海涛，等. 一种基于“是一个”模式的下位概念获取方法. 计算机科学，2006，33(9)：146～151

15 Tian Guogang, Cao Cungen, Liu Lei, et al. MFC: A Method of Co-referent Relation Acquistion from Large-scale Chinese Corpora [C]. In: Fuzzy Systems and Knowledge Discovery(FSKD 2006), Xi'an, China, 2006. 1259～1268