一种基于概率粗糙集模型的增量式规则学习算法*)

付长龙[1,2] 杜旭辉[2] 姚全珠[2]

(西北工业大学自动化学院 西安 710072)[1] (西安理工大学计算机科学与工程学院 西安 710048)[2]

摘 要 提出了一种基于概率粗糙集模型的增量式规则学习算法。该算法能够有效地从不一致和含有噪声的决策表中提取带有确定性因子和支持数的决策规则,并且所提取出的规则具有很好的抗噪声能力。同时,算法的动态调整策略可以满足规则的动态更新。最后将该算法应用于一个实例分析中,提取了满足给定参数的决策规则,分析结果验证了该算法在规则提取中的合理性。

关键词 粗糙集,概率粗糙集,规则学习,不一致决策表

An Incvementral Rules Learning Algorithm Based on Probabilistic Rough Set Model

FU Chang-long[1,2] DU Xu-hui[2] YAO Quan-zhu[2]

(School of Automation, Northwestern Polytechnical University, Xi'an 710072, China)[1]

(School of Computer Science, Xi'an University of Technology, Xi'an 710048, China)[2]

Abstract In this paper, an incremental rule learning algorithm based on probabilistic rough set model is presented. The method can extract rules with certainty factor and supporting number from inconsistent decision table effectively, and the rules extracted by the method have a good noise resisting ability. At the same time, the dynamic adjustment mechanism of the algorithm can satisfy the dynamic renewal of rules. At the end of the paper, the rules which meet the specified criteria are derived by applying the method to an example, and the result validated the rationality of the method in rules extraction.

Keywords Rough set, Probabilistic rough set, Rules learning, Inconsistent decision table

1 引言

粗糙集理论是一种处理不精确、不确定和不完整数据的数学理论,它已经成功地应用到了机器学习、知识发现、决策分析、模式识别等领域[1-3]。传统的 Pawlak[4] 粗糙集理论是基于可利用信息的完整性的,它所处理的分类问题是完全正确的或肯定的。由于它是严格按照等价关系来分类的,因此它对对象的分类结果只有"属于"和"不属于",而没有某种程度上的"属于"或"包含"。这种形式不能够识别非决策关系,如可能以小于 1 的概率推导出预测规则,目前在这方面已经做出了不少研究[5-7]。然而,在实际应用中,知识库中的数据往往是由随机原因或统计得到的,即知识库中的数据很可能存在噪声和某种程度的不一致,从而导致分类模式的交迭,不能够产生强的决策规则,但可以导出用于概率估计的非强决策规则。

本文对概率粗糙集[8,9]进行了深入的研究,在此基础上提出了一种基于概率粗糙集模型的带有决定性因子和支持数的规则学习算法。该算法很好地解决了从不一致决策表和含有噪声的数据中提取规则的问题,通过该算法提取的规则能够对概念有一个更完整的刻画。

2 概率粗糙集模型

定义 1 设 U 是有限对象构成的论域,R 是 U 上的等价关系,其构成的等价类为 $U/R=\{X_1,X_2,\cdots,X_n\}$,记 x 所在的等价类为 $[x]$,令 P 为定义在 U 的子集类构成的 σ 代数上的概率测度,三元组 $A_p=(U,R,P)$ 称为概率近似空间。U 中的每个子集称为概念,它代表了具有一定概率的随机事件。$P(X|Y)$ 表示在事件 Y 发生的条件下 X 出现的概率,也可以解释为随机选择的对象在概念 Y 的描述下属于 X 的概率。设 $0\leqslant\beta<\alpha\leqslant1$,对于任意 $X\subseteq U$,我们定义 X 关于概率近似空间 $A_p=(U,R,P)$ 依参数 α,β 的下近似 $\underline{PA}_\alpha(X)$ 和上近似 $\overline{PA}_\beta$ 如下:

$$\underline{PA}_\alpha(X)=\{x\in U|P(X|[x]_R)\geqslant\alpha\}$$

$$\overline{PA}_\beta(X)=\{x\in U|P(X|[x]_R)>\beta\}$$

X 关于 A_P 依参数 α,β 的概率型正域、边界和负域分别为

$$POS(X,\alpha)=\underline{PA}_\alpha(X)=\{x\in U|P(X|[x]_R)\geqslant\alpha\}$$

$$BN(X,\alpha,\beta)=\{x\in U|\beta<P(X|[x]_R)<\alpha\}$$

$$NEG(X,\beta)=U-\overline{PA}_\beta=\{x\in U|P(X|[x]_R)\leqslant\beta\}$$

由此可见,当一个对象属于 $POS(X,\alpha)$ 或 $NEG(X,\beta)$ 时,可以从概率的意义上确定这个对象满足概念 Y 的程度,但不能确定边界中研究对象是否匹配概念 Y。

定义 2 对于概率近似空间 (U,R,P),系统的不确定性用系统的熵 $H(R^*)$ 来表示,其计算公式如下:

$$H(R^*)=-\sum_{i=1}^{n}P(X_i)\log P(X_i)$$

其中 $R^*=U/R=\{X_1,X_2,\cdots,X_n\}$ 为 R 在 U 上导出的划分,系统的熵常称为系统的初始熵。

*)国家自然科学基金资助项目(编号:50279041)、国家"863"高技术研究发展计划资助项目(编号:2005AA113150)。付长龙 博士研究生,副教授,主要研究方向为数据库技术与网络技术;杜旭辉 硕士研究生,主要研究方向为网络安全技术;姚全珠 教授,博士,主要研究方向为软件工程方法与网络技术。

定义 3 设 $K=(U,A,V,\rho)$为一个信息系统，$R^*=U/R=\{X_1,X_2,\cdots,X_n\}$为等价关系 R 在 U 上导出的划分，S 为 U 上另一等价关系，且 $S^*=\{Y_1,Y_2,\cdots,Y_m\}$，则在已知知识 R^* 时知识 S^* 的正则条件熵定义为

$$H_0(S^*|R^*)=-\sum_{i=1}^{n}P(X_i)\sum_{j=1}^{m}P(Y_j|X_i)\log P(Y_j|X_i)/\log m$$

显然 $0\leqslant H_0(S^*|R^*)\leqslant 1$。正则条件熵表示系统中属性 S 基于属性 R 的信息依赖性，常记为：$R\xrightarrow{H_0(S^*|R^*)}S$。特别地，若 $H_0(S^*|R^*)=0$，则划分 S^* 称为完全信息依赖于划分 R^*。若 $H_0(S^*|R^*)=1$，则划分 S^* 称为完全信息独立于划分 R^*。

定义 4 若存在 $N\subset M\subset A$(A 为一个属性集)，使得 $H_0(M^*|N^*)=H_0(M^*|M^*)$，则称 M 为统计依赖的，否则称 M 是统计独立的。若 N 是 M 的一个极大统计独立子集，则 N 称为 M 的一个统计约简，M 的所有统计约简记为 $sred(M)$。M 的所有统计约简的交集称为 M 的核，记为 $score(M)$。

定义 5 若存在 M 的一个真子集 L 使 $H_0(B^*|M^*)=H_0(B^*|L^*)$，则属性子集 $M\subseteq A$ 称为关于属性集 $B\subseteq A$ 是统计依赖的，否则称 M 关于 B 是统计独立的。若 N 是 M 关于 B 的一个极大统计子集，则 N 称为 M 关于 B 的一个相对约简，M 关于 B 的所有相对约简记为 $sred_B(M)$，所有这些约简的交集称为 M 关于 B 的相对核，记为 $score_B(M)$。

定义 6 若 $H_0(C^*|(C-\{a\})^*)=H_0(C^*|C^*)$，则称 a 在 C 中是统计可省略的，否则为 C 中统计不可省略的。

3 基于概率粗糙集的规则学习算法

在粗糙集规则提取方面，已经做了很多相关的研究，但仍然存在不少问题。例如，文献[3]中的方法不能有效地从含有噪声的信息中提取规则；文献[7]所提出的方法不能对规则进行后继的动态更新，因此在实际应用中存在很大的缺陷。

针对上述问题，本文提出了一种基于概率粗糙集的增量式规则学习算法，算法分为规则学习和动态更新两个阶段。

第一阶段：规则学习阶段

Step1：对数据进行离散化处理，常用的离散化方法有等间距法、等频距法、信息熵法等，本文采用信息熵法[10]对连续属性进行离散化处理。

Step2：设离散化后的数据对象构成的论域为 U，属性集 $A=C\cup D$，且 $C\cap D=\Phi$，其中 $C=\{c_1,c_2,\cdots,c_m\}$ 为条件属性集，$D=\{d_1,d_2,\cdots,d_n\}$ 为决策属性集。

Step3：计算每一个条件属性 $c_i(i=1,2,\cdots,m)$ 对论域 U 的划分 U/c_i，进一步求得 $C^*=U/ind(C)$。

Step4：计算每一个决策属性 $d_j(j=1,2,\cdots,n)$ 对论域 U 的划分 U/d_j，进一步求得 $D^*=U/ind(D)$。

Step5：根据正则条件熵计算公式，计算已知知识 C^* 时知识 D^* 的正则条件熵 $H_0(D^*|C^*)$。

Step6：计算 $H_0(D^*|\{C-c_i\}^*)$，判断它的值是否等于 $H_0(D^*|C^*)$。

Step7：若不相等，则说明属性 c_i 关于决策属性集 D 是必须的；如果相等，则属性 c_i 关于决策属性集 D 是多余的，使 $C=C-\{c_i\}$。

Step8：使 $i=i+1$，如果 $i\geqslant m$，转到 *step*9，否则转到 step5；

Step9：C 中剩余的属性就是条件属性的统计约简，从决策表中删除 C 中不包含的属性列，得到决策表的约简。

Step10：给决策表中的对象进行编号，重复出现的对象使用相同的编号 i，并统计对象 i 在决策表中重复出现的次数 p_i，用 p_i 表示决策表中对象 i 的支持数。

Step11：删除决策表中重复出现的对象，得到约简后的决策表，设其条件属性集为 C'，决策属性集为 D。

Step12：引入参数 α 和 β(其中 $0\leqslant\beta<\alpha\leqslant 1$)，对于约简后的决策表，根据满足条件属性 C' 的描述 $des(C')$ 的研究对象是否在一定程度上(用确定性因子 c 刻画)满足决策属性 D 的描述 $des(D)$来抽取带确定性因子 c 和支持数 p 的决策规则，其具体规则为：

(1)如果 $P(D|C')>\alpha$，则 $des(C')\xrightarrow{(c,p)}des(D)$；

(2)如果 $P(D|C')<\beta$，则 $des(C')\xrightarrow{(c,p)}\neg des(D)$；

(3)如果 $\alpha\leqslant P(D|C')\leqslant\beta$，则 $des(C')\xrightarrow{(c,p)}$unknown。

在这种情况下，从现有的知识不能形成决策规则。在规则提取过程中，对于 $des(C')$相同，而 $des(D)$不同的对象，取支持数最大的对象的 $des(D)$为规则的 $des(D)$；规则中的确定性因子定义为：

$$c=\max\{P(D|C'),1-P(D|C')\}$$

第二阶段：规则动态更新阶段

Step1：根据已有规则库中的规则对新加入的数据对象 R 进行匹配，匹配结果可以分为以下 3 种情况：

(1) 如果 R 的条件属性和决策属性都与某已有的规则相匹配，则转到 step2；

(2) 如果 R 的条件属性和已有规则匹配，而决策属性不匹配，则转到 step3；

(3) 如果 R 的条件属性和已知规则不匹配，则转到 step4。

Step2：对已有规则的确定性因子 c 和支持数 p 按照如下方法进行调整：规则中新的确定性因子 $c'=\frac{(p+p_R)c}{p+cp_R}$，支持数调整为 $p'=p+p_R$，其中 p_R 为数据对象 R 的支持数。

Step3：比较 R 的支持数 p_R 和现有规则的支持数 p 的大小，并按以下情况处理：

(1)若 $p_R\geqslant p$，则规则新的确定性因子 $c'=\frac{cp_R}{p+cp_R}$，新的支持数为 $p'=p_R$，规则的决策属性部分替换为 R 属性部分的描述 $des(D_R)$；

(2)若 $p_R<p$，规则中新的确定性因子 $c'=\frac{(p-p_R)c}{p-cp_R}$，新的支持数为 $p'=p-p_R$；

判断新的确定性因子 c' 与参数 α 和 β 的关系：如果 $c'>\alpha$，则原规则的条件属性和决策属性保持不变；如果 $c'<\beta$，则原规则的条件属性保持不变，决策属性的描述由 $des(D)$变为 $\neg des(D)$；如果 $\beta\leqslant c'\leqslant\alpha$，则决策属性部分变为 unknown，说明由现有的知识无法做出决策，从规则库中删除该条规则。

Step4：按照 R 的条件属性和决策属性生成新的规则，并加入到已有的规则库中。

4 实例分析

某经离散化的销售量信息决策表如表 1 所示。表中的属

性含义分别为：c_1 为商品质量，c_2 为商品价格，c_3 为促销，c_4 为服务，d 为商品销售量。

表 1　销售量信息决策表

U	条件属性(C)				决策属性(d)
	c1	c2	c3	c4	
1	好	低	无	好	高
2	好	低	有	好	高
3	好	低	无	好	高
4	好	低	无	好	高
5	好	低	无	好	高
6	好	高	有	差	中
7	好	高	无	差	低
8	差	高	有	好	低
9	差	高	有	好	中
10	差	低	有	好	中
11	差	低	无	差	低
12	好	低	无	好	中
13	好	低	无	好	高

(1)由于该表已经经过了离散化，因此可以直接求每个条件属性对论域 U 的划分如下：

$U/c_1=\{\{1,2,3,4,5,6,7,12,13\},\{8,9,10,11\}\}$

$U/c_2=\{\{1,2,3,4,5,10,11,12,13\},\{6,7,8,9\}\}$

$U/c_3=\{\{1,3,4,5,7,11,12,13\},\{2,6,8,9,10\}\}$

$U/c_4=\{\{1,2,3,4,5,8,9,10,12,13\},\{6,7,11\}\}$

由上面的划分结果可以计算出：

$C^*=U/ind(C)=\{\{1,3,4,5,12,13\},\{2\},\{6\},\{7\},\{8,9\},\{10\},\{11\}\}$

(2)计算决策属性对论域 U 的划分为：

$D^*=U/ind(D)=U/d=\{\{1,2,3,4,5,13\},\{6,9,10,12\},\{7,8,11\}\}$

(3)由上面的计算结果可以计算出已知知识 C^* 时，知识 D^* 的正则熵为：

$$H_0(D^*\mid C^*)=-\sum_{i=1}^{7}P(X_i)\sum_{j=1}^{3}P(Y_j\mid X_i)\log P(Y_j\mid X_i)/\log 3=0.286$$

(4)同样根据正则条件熵的定义，通过计算可以判断 $H_0(D^*\mid(C-c_i)^*)$ 和 $H_0(D^*\mid C^*)$ 的关系分别为：

$H_0(D^*\mid(C-c_1)^*)-H_0(D^*\mid C^*)=0$

$H_0(D^*\mid(C-c_2)^*)-H_0(D^*\mid C^*)\neq 0$

$H_0(D^*\mid(C-c_3)^*)-H_0(D^*\mid C^*)\neq 0$

$H_0(D^*\mid(C-c_4)^*)-H_0(D^*\mid C^*)=0$

由计算结果可以知道在条件属性集 C 中，属性 c_1 或 c_4 是多余的，也就是说，从 C 去掉 c_1 或 c_4 是不会影响决策表中条件属性和决策属性的总体依赖性的。

(5)假设我们从 C 中去掉属性 c_4，对约简后的条件属性集 $C'=C-\{c_4\}$ 重复上述步骤，可以得到以下结论：

$H_0(D^*\mid(C'-c_1)^*)-H_0(D^*\mid C'^*)\neq 0$

$H_0(D^*\mid(C'-c_2)^*)-H_0(D^*\mid C'^*)\neq 0$

$H_0(D^*\mid(C'-c_3)^*)-H_0(D^*\mid C'^*)\neq 0$

这说明 C' 中每个属性都是不可省略的，也就是说 C' 就是 C 关于决策属性集 D 的一个最小统计约简。

(6)从决策表中删除属性 c_4，根据算法 step9-step11 可以得到约简后的决策表，如表 2 所示。

表 2　约简后的决策表

U	条件属性(C)			决策属性	支持数
	c1	c2	c3	(d)	
1	好	低	无	高	5
2	好	低	无	中	1
3	好	低	有	高	1
4	好	高	有	中	1
5	好	高	无	低	1
6	差	高	有	低	1
7	差	高	有	中	1
8	差	低	有	中	1
9	差	低	无	低	1

(7)假设我们引入参数 $\alpha=0.6$，$\beta=0.36$，根据算法的 step12，可以从决策表中提取带确定性因子 c 和支持数 p 的规则如下：

$c_1='好'\wedge c_2='低'\wedge c_3='无'\xrightarrow{(83.3\%,5)}d='高'$

$c_1='好'\wedge c_2='低'\wedge c_3='有'\xrightarrow{(100\%,1)}d='高'$

$c_1='好'\wedge c_2='高'\wedge c_3='有'\xrightarrow{(100\%,1)}d='中'$

$c_1='好'\wedge c_2='高'\wedge c_3='无'\xrightarrow{(100\%,1)}d='低'$

$c_1='差'\wedge c_2='低'\wedge c_3='有'\xrightarrow{(100\%,1)}d='中'$

$c_1='差'\wedge c_2='低'\wedge c_3='无'\xrightarrow{(100\%,1)}d='低'$

(8)假设在前面决策系统的基础上加入一个新的决策表，该决策表按照上述过程得到约简后的决策表如表 3 所示。

表 3　约简后的决策表

U	条件属性(C)			决策属性	支持数
	c1	c2	c3	(d)	
1	好	低	无	高	1
2	好	低	有	中	3
3	差	高	有	低	1

因此，按照规则的动态调整算法，可以得到调整后的决策规则如下：

$c_1='好'\wedge c_2='低'\wedge c_3='无'\xrightarrow{(85.7\%,6)}d='高'$

$c_1='好'\wedge c_2='低'\wedge c_3='有'\xrightarrow{(75\%,2)}d='中'$

$c_1='好'\wedge c_2='高'\wedge c_3='有'\xrightarrow{(100\%,1)}d='中'$

$c_1='好'\wedge c_2='高'\wedge c_3='无'\xrightarrow{(100\%,1)}d='低'$

$c_1='差'\wedge c_2='低'\wedge c_3='有'\xrightarrow{(100\%,1)}d='中'$

$c_1='差'\wedge c_2='低'\wedge c_3='无'\xrightarrow{(100\%,1)}d='低'$

$c_1='差'\wedge c_2='高'\wedge c_3='有'\xrightarrow{(100\%,1)}d='低'$

结束语　Pawlak 粗糙集模型是基于可利用信息的完整性的，它忽视了可利用信息的完整性和可能存在的统计信息，因此这类模型对于不协调或存在噪声数据的决策表的规则提取就显得无能为力。针对上述问题，本文提出了一种基于概率粗糙集的增量式规则学习算法，该算法能够有效地从不一致和含有噪声的决策表中提取带有确定性因子和支持数的决策规则。规则提取算法的动态更新策略可以满足规则的动态调整，从而可以使提取的规则更加合理。另外，在规则提取时

可以通过设定参数 α 和 β 来控制所提取决策规则的强弱性，当 $\alpha=1$ 且 $\beta\neq0$ 时，所提取的规则就是完全确定的强决策规则，这时规则的容错性能就比较差。

最后，通过实例分析对算法的有效性进行了验证。从销售量信息决策表中提取的规则可以看出，该方法所提取出的规则是比较合理的，所提取出的规则可以很好地对不一致信息进行分类。

参考文献

[1] Pawlak Z. Rough classification[J]. International Journal of Man-machine Studies, 1984, 20:469-483

[2] Pawlak Z. Rough sets and Decision Analysis[C]. Fifth IIASA Workshop on Decision Analysis and Support, 1998:123-127

[3] 印勇，曹长修，张邦礼. 基于粗糙集理论的分类规则发现[J]. 重庆大学学报，2000，23(1)：63-65

[4] Pawlak Z. Rough sets. International Journal of Computer and information Science, 1982, 11: 341-356

[5] Kryszkiewicz M. Rules in incomplete information system[J]. Information Sciences, 1999, 113:271-292

[6] Kryszkiewicz M. Rough set approach to incomplete information system. Information Sciences[J], 1998, 112: 39-49

[7] 黄兵，魏大宽，周献中. 不完备信息系统最大分布约简及规则提取算法[J]. 计算机科学，2004，32(8A)：53-56

[8] 张文修，吴伟志，梁吉业，等. 粗糙集理论与方法. 北京：科学出版社，2001

[9] 王基一，许黎明. 概率粗糙集模型[J]. 计算机科学，2002，28(8A)：76-78

[10] 谢宏，程浩忠，牛东小. 基于信息熵的粗糙集连续属性离散化算法[J]. 计算机学报，2005，28(9)：1570-1574

(上接第 124 页)

(2)-(4)得到公式(5)。

$$\mathrm{Sim}(A,B)=w_{name}\,\mathrm{Sim}_{name}(A,B)+w_{attribute}\,\mathrm{Sim}_{attribute}(A,B)+w_{relation}\,\mathrm{Sim}_{relation}(A,B) \quad (8)$$

其中，w_{name}，$w_{attribute}$，$w_{relation}$ 分别为概念名称相似性、概念属性集合相似性、相关概念相似性的权重，$w_{name}+w_{attribute}+w_{relation}=1$，$w_{name}\neq0$。

4 实验及分析

为了对以上方法进行评估，作者在两个本体数据上作了实验(表 1)。它们都描述了 Cornell 大学和 Washington 大学的课程体系。

表 1

	Ontologies	Concepts	Number of instances
Course Catalog Ⅰ	Cornell	34	1526
	Washington	39	1912
Course Catalog Ⅱ	Cornell	176	4360
	Washington	166	6975

根据公式(2)-(4)可以得出表 2。

表 2

Ontologies	sim_{name}	$\mathrm{sim}_{attribute}$	$\mathrm{sim}_{relation}$
Course Catalog Ⅰ Course Catalog Ⅱ	0.46	0.92	0.87

概念相似度的综合计算：

本文把概念相似度看成基于名称、属性和关系相似度计算的综合，分别取权值 $w_{name}=0.4$，$w_{attribute}=w_{relation}=0.3$。由综合公式(5)得：

$$\mathrm{Sim}(C,W)=0.4\times0.46+0.3\times0.92+0.3\times0.87=0.721$$

对表(1)中的数据我们用 ACAOM[2] 方法进行计算可得到。

$$\mathrm{Sim}(C,W)=0.5\times0.46+0.5\times0.96=0.710$$

时间复杂度分析：

影响复杂度的关键因素有生成候选映射空间的复杂度、候选映射对的个数、相似度度量的复杂程度。如果本体中的实体个数为 n，在不做任何改善的情况下，可能的候选映射对的个数为 $n*n$。为了提高映射效率，对候选映射对的选择采用以下策略：对基距离排序，选择基距离较小的作为候选映射对。所以生成后选概念集在最坏情况下所需时间：$T_{candidate}=O(n\log(n))$，在候选概念集中，对每个概念来说，原本有 n 个概念要与之比较，现在只固定 k 个，所以比较的概念 $N_{compare}$ 最多为 $n*k$ 个，所有相似度计算的最大复杂度为 $T_{compare}=m*O(\log^2 n)$，而对相似度的综合时复杂度为 $T_{integrate}=O(n)$。在使用候选概念集和信息增益的情况下为：$O(n\log(n))+O(n)*m*O(n*n)+O(n)=O(n\log(n))$，在直接使用所有相关概念的情况下为：$O(n\log(n))+O(n)*m*O(\log^2(n))+O(n)=O(n\log^2(n))$。

结束语 针对本体映射中概念相似度计算量大、计算精度不高等问题，本文给出了一种基于语义 Web 的本体概念相似度的计算方法，通过给定一定权值提高概念相似度的精确度，并引入信息增益和候选概念集。从相似度计算结果可得出该方法的精度比 ACAOM 方法高，从时间复杂度分析可以得出其复杂度要比 ACAOM 方法的复杂度小。

参考文献

[1] Doan A, Madhavan J Y, Domingos P, et al. Learning to map between ontologies on the semantic web// Proceedings of the World Wide Web Conference. 2002: 662-673

[2] 黄烟波，张红宇，李建华，等. 本体映射方法研究. 计算机工程与应用，2005：27-33

[3] 邓志鸿，唐世渭，张铭，等. Ontology 研究综述. 北京大学学报，2002(5)：730-738

[4] Zhou Chunguang, Wang Ying. A Composite Approach for Ontology Mapping//Proceeding of the 23rd British Database Conference. 2006

[5] Sekine S, Sudo S, Ogino S. Statistical Matching of two ontologies//Proceedings of the SIGLEX99: Standerdizing Lexical Resources. Maryland, USA, 1999: 69-73

[6] Pedersen T, Patwardhan S, Michelizzi S. WordNet: Similarity - Measuring the Relatedness of Concepts//Proceedings of Fifth Annual Meeting of the North American Chapter of the Association for Computational Linguistics[C]. [s. l.], 2004

[7] Qian Peng-fei, Zhang Shensheng. Ontology Mapping Approach Based on Concept Partial. IEEE, 2006(7): 4107-4112

[8] Ehrig M, Sure Y. Ontology Mapping-An Integrated Approach. ESWS, 2004: 76-91

[9] 钟宁，尹旭日，陈世福. 基于信息增益的最佳属集性发现方法. 小型微型计算机系统，2002(4)：444-446