基于互信息和本体的协同检索模型的研究*)

周竹荣　邱玉辉　夏　磊

（西南大学计算机与信息科学学院　重庆 400715）

摘　要　在信息检索中，用户习惯用尽可能少的关键字来检索信息，这必然会导致检索结果与用户需求存在较大偏差。针对这一问题，我们提出了基于互信息的语义扩展模型(QSE_BMI)[2]，结合用户兴趣模型，对用户输入的查询问句进行语义扩展。本文在 QSE_BMI 基础上，利用互信息与本体互补性，建立基于互信息和本体的协同检索模型，从而提高了信息检索的查全率与查准率。

关键词　互信息，本体，互补，语义发现，领域

Research on the Mutual Information and Ontology Based Collaboration Retrieve Model

ZHOU Zhu-Rong　QIU Yu-Hui　XIA Lei

(School of Computer & Information Science, Southwest University, Chongqing 400715)

Abstract　People are used to retrieve information they need with keywords as few as possible. It obviously causes the results they get may be different with their willing. According to this phenomenon, we propose the QSE_BMI model[2] to expand the query which user inputs based on user's interesting model. We make use of the complement between the mutual information and ontology to propose the mutual information and ontology based collaboration retrieve model for the target of improving the chance of recall and precision.

Keywords　Mutual information, Ontology, Complement, Semantic-finding, Domain

在信息检索的过程中，通过提取用户问句中的关键字得到的查询结果与用户的查询意图常有偏差[1]，这就需要信息检索系统对用户问句的关键字进行语义扩展。根据对文[2]中提出的 QSE_BMI 模型进行了深入的研究之后发现，利用互信息对领域文档集进行词对之间互信息量计算得到互信息矩阵时，会出现较大的冗余互信息，这些冗余互信息来源于：1)冗余相关词对；2)不相关词对。这就给基于互信息的语义扩展带来了较大的冗余度，影响了最终检索的查准率。

基于以上的原因，我们把领域本体与互信息相结合，利用领域本体对互信息矩阵中冗余互信息进行过滤，利用互信息矩阵对领域本体进行语义补充。我们发现互信息矩阵，当相关词对出现的概率较高时，则该关键词对存在可能的语义关系。所以我们提出语义阈值概念，当关键词对出现的概率高于该阈值，且不在领域本体中出现，则对领域本体进行语义补充，从而形成领域本体与互信息互补。

1　问题提出

在文[2]中我们提出的 QSE_BMI 模型取得了较好的效果。但随着训练文档集的增加，互信息矩阵出现较大无关信息。例如，在"信息技术"领域中，我们选取训练文档集的文档数量分别为：100，200，500，1000，得到图 1 的冗余互信息变化曲线。随着训练文档集的数量的逐渐增加达到了某一拐点之后，在互信息矩阵当中出现的冗余互信息数量也是成几何数增加。

出现该问题的原因在于，互信息计算的是两个随机变量之间相关联的程度[2]。因此对于文档集当中的词对而言，虽然我们是在一定窗口(即抽取关键字的范围，比如文件中相邻的十个词)当中取出的两个关键字，但是在这个窗口当中，同样会出现两个无关的关键字之间有着互信息量的存在。因此，这就使得 QSE_BMI 模型在运行时，出现了大量的冗余甚至是错误的词对之间的互信息量，从而为我们的语义扩展带来了较大的干扰。所以，我们需要一种方法，来过滤互信息中的冗余互信息。

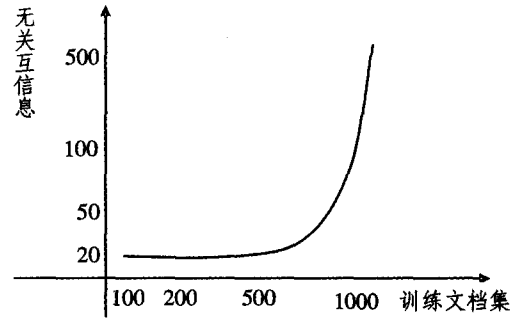

图 1　冗余互信息曲线图

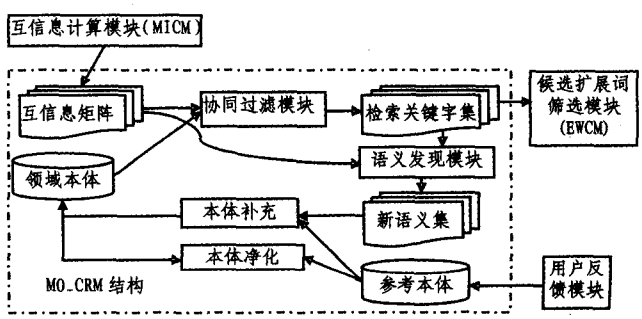

图 2　MO_CRM 模型总体结构图

*)本文受到国家 973 重大基础研究项目基金资助(项目编号为：2003CB317008)。**周竹荣**　博士生，主要研究方向：ICAI，远程教育；**邱玉辉**　教授，博士生导师，主要研究方向：人工智能、机器学习；**夏　磊**　硕士研究生，主要研究方向：智能计算机辅助教育。

经过研究，我们发现本体与互信息矩阵具有互补性。首先，互信息矩阵具有动态性和冗余互信息这两个主要的特点。其次，本体具有静态性与更新慢的特点[4]。从这两点我们可以经分析得到，本体的弱点却恰好是互信息的长处。于是我们提出，用本体与互信息相结合，利用其互补的优势，建立一个基于互信息和本体的协同检索模型(MO_CRM)，来解决以上的问题。

2 MO_CRM 模型

在 MO_CRM 模型中，本体与互信息互为补充，互相利用：

1)利用本体对互信息矩阵过滤

根据本体当中所定义的概念之间的关系，或是关键词之间的关系，对互信息矩阵进行过滤。以去除掉互信息矩阵当中一些不符合正确语义关系的词对。

2)利用互信息对本体进行语义补充

利用互信息的动态性和实时性，发现语言词汇的一些新的搭配使用方法，对本体进行完善和补充，实现本体的更新和学习。

2.1 MO_CRM 模型总体结构

经过对 QSE_BMI 的改造，引入本体，我们得到 MO_CRM 模型，如图 2 所示。

2.1.1 相关定义

为讨论方便，我们给出以下定义：

定义 1(领域本体 Domain Ontology) 领域本体指的是研究与一个特定领域相关的术语或词汇，专注于解决领域知识的抽象，较为具体。本文所用的领域本体，主要存放文档集中各词汇之间的关系[5]。

定义 2(参考本体:Reference Ontology) Guarino 曾经提出以详细程度对 Ontology 进行分类[6]，详细程度高的本体称为参考本体。本文中的参考本体主要指在由用户反馈模块得到的用户对检索结果中错误的结果所组成的本体。

定义 3(语义发现 Semantic Discovery) 语义发现指的是对关键词对的新的搭配使用方法的发现。本文当中利用关键词对之间的互信息量与阈值之间的关系，来进行语义的发现。

定义 4(本体净化 Ontology Purify) 本体净化是将本体当中错误的语义关系进行更改或删除，以达到净化本体，以便本体达到更为准确的效果[7]。

定义 5(本体补充 Ontology Complementarity) 本体补充指的是将通过语义发现所得到的新的正确的词对搭配关系对本体自身进行补充，进行完善，使本体与语言使用规则保持一致性。

2.1.2 模型概述

根据以上给出的模型的总体框架图，对模型进行了如下的描述：

1) 由领域专家建立领域本体。

2) 协同过滤模块在领域本体的帮助下对互信息矩阵进行处理，过滤掉领域本体中不存在，而互信息矩阵存在的词对，得到关键字集。

3) 语义发现模块对关键字集进行分析，发现新的语义词对，得到新语义集。

4) 本体补充模块将新语义集加入领域本体。

5) 本体净化模块在参考本体的帮助下，将本体当中错误的语义关系进行更改或删除。

在模型中，要解决二个关键问题：

1) 如何从互信息矩阵中发现新的语义信息

互信息矩阵存放的是通过对训练文档集进行互信息计算之后，所得到的关键词对之间的相关联的度量。但是这种度量只是一个量化的过程，并不能保证矩阵当中具有互信息量的两个关键词是不是真的就具有语义关联。因此，如何从互信息矩阵中发现出新的语义信息，就是我们要解决的首要问题。

2) 如何净化本体的错误语义信息

本体虽然具有准确描述语义关系的特别，但由于语言自身是发展的，因此本体在用于描述语言的时候就会出现过时或是偏差的情况。于是我们要对这种情况进行清理，达到净化本体的目的。这种语言的新的使用方法，我们只能通过用户所反馈回来的信息来加以识别并处理，将这些错误的关系从原领域本体当中删掉，从而实现本体的净化的功能。

2.2 形式化描述

为了更好地对上述模型进行描述，我们对模型当中的各组件进行了相应的形式化描述。

1. 互信息矩阵的形式化描述如下：

首先定义关键词向量：$K=(key_0, key_1, key_2, \cdots, key_n)$

互信息矩阵可定义为：$MIM=(M_0, M_1, \cdots, M_n)$

其中：$M_i=(MI_{i0}, MI_{i1}, \cdots, MI_{im})$ 表示的是两个关键字之间的互信息量。MI_{ij} 计算公式如下所示：

$$MI_{ij}=MI(KEY_i, KEY_j)$$

$$=I(key_i, key_j)=\log\left(\frac{p(key_i, key_j)}{p(key_i)\,p(key_j)}\right)$$

$$=\log\frac{\frac{c(key_i, key_j)}{N}}{\frac{c(key_i)}{N}\times\frac{c(key_j)}{N}}=\log\frac{c(key_i, key_j)\times N}{c(key_i)\times c(key_j)} \quad (1)$$

根据最大似然估计，在语料规模足够大的情况下，可以认为单词出现的概率为其出现的次数。其中 $c(key_i, key_j)$ 表示单词 key_i、key_j 有序同现的次数，$c(key_i)$ 表示关键词 key_i 出现的次数，$c(key_j)$ 表示关键词 key_j 出现的次数，N 为文档库中所有单词的个数。

2. 领域本体的形式化描述如下：

$DO=(C, R, A, I, P)$

其中：C 表示的是类(classes)或概念(concepts)：指任何事务，如工作描述、功能、行为、策略和推理过程。从语义上讲，它表示的是对象的集合，其定义一般采用框架(frame)结构，包括概念的名称，与其他概念之间的关系的集合，以及用自然语言对概念的描述。

$R=\{$part-of, kind-of, instance-of, attribute-of$\}$，表示的是词对之间的关系(relations)。

A 表示的是公理(axioms)：代表永真断言，如概念乙属于概念甲的范围。

I 表示的是实例(instances)：代表元素。从语义上讲实例表示的就是对象。

P 表示的是属性(Property)：代表对概念或性质的描述。.

3. 关键字集的形式化描述：$TermSet=(Term_1, Term_2, Value)$

其中：$Term_1=\{Term_{11}, Term_{12}, \cdots, Term_{1n}\}$ 表示的是经过协

同过滤模块处理之后所得到的关键字/词。

$Term_2=\{Term_{21},Term_{22},\cdots,Term_{2n}\}$表示的是经过协同过滤模块处理之后所得到的与 $Term_1$ 集当中所对应的关键字/词。

$Value=\{m|0<m<1\}$ 表示的是 $Term_1$ 与 $Term_2$ 当中相对应的词之间的互信息量。

4、参考本体的形式化描述：$RefO=(\{NewKey, FormKey, Flag\})$

其中：$NewKey=\{Key_1,Key_2,\cdots,Key_N\}$ 表示的是用户反馈的关键字。此类关键字分为两类，一类是在领域本体当中没有的，另一类是在领域本体当中存在错误关系的关键字。

$FormKey=\{Key_1',Key_2',\cdots,Key_N'\}$ 表示的是与用户反馈的关键字相关联的且已存在于领域本体当中的关键字。

Flag={0,1} 用于表示参考本体当中的信息是不是存在新的语义，是则置 1，否则置 0。Flag 值的默认值为 0。

5、新语义集的形式化描述：$NewSemanticSet=(Find_i, Find_j)$

与参考本体类似，新语义集也用相同的方法进行形式化描述如下：

$Find_i=\{Find_{i1},Find_{i2},\cdots,Find_{iK}\}$ 表示的是发现的可能具有新语义的关键字/词。

$Find_j=\{Find_{j1},Find_{j2},\cdots,Find_{jK}\}$ 表示的是与 $Find_i$ 相对应的具有语义关系的关键字/词。

3 算法描述

通过以上介绍，我们提出的 MO_RCM 模型中有四个主要的模块。其中每一个模块都对应着一个算法。由于篇幅所限，本文中我们仅列出其中的协同过滤算法与语义发现算法。至于本体净化算法与本体补充算法，我们将另文陈述。

3.1 协同过滤算法(Collaboration Filtering Algorithm, CFA)

在经过了文[2]中我们提出的 QSE_BMI 模型当中的“互信息计算模块(MICM)”处理之后，便得到了对文档集的互信息矩阵(MIM)。根据本文提出的 MO_CRM 模型，我们将结合专家制作的领域本体对互信息矩阵进行过滤。将一些不存在于领域本体当中的具有语义的关键字滤去。具体算法如下：

Input：互信息矩阵 MIM，领域本体 DO；
Output：关键字集 $TermSet$；
Procedure：
Step1：for 互信息矩阵中的词对 $MIM_{ij}(i,j=0\text{ to }n)$
If ($MIM_{ij}>\alpha$)
$PotentialSet_{ij}=MIM_{ij}$；其中 $PotentialSet$ 为潜在语义关系集，α 为初级阈值，用于选出具有潜在语义关系的关键词对，取值范围为(1>α>0)。
Step2：集合 $MSet=PotentialSet_{ij}\cap DO_{ij}$；
Step3：如果 $MSet\neq\phi$，则将 $TermSet\leftarrow MSet$，否则 $TermSet=\phi$。

3.2 语义发现算法(Semantic Discovery Algorithm, SDA)

在前文我们已经阐述过，虽然领域本体是由领域专家所创建的，具有准确的语义描述能力，但是，语言的使用是一个发展的过程。随着时间的推移，人们对语言的把握多多少少会有所变化。但领域本体，一旦建立，要改变或是修改相当耗时，耗力，而且这种变化还不易于人为的发现和总结。因此，我们有必要建立一个语义发现模块，来对新的语言使用方法进行获取，从而为领域本体的更新做准备。

具体的语义发现算法(SDA)，算法如下：

Step1：集合 Q=*TermSet*−*PotentialSet*。其中 TermSet 与 PotentialSet 均为在 CFA 算法中得到的相关集合。

Step2：If (Q =ϕ)
集合 NewSemanticSet=ϕ；
Return；
Else Setp3；

Step3：if (集合 Q 中的词对的互信息量 $Q_{ij}>\beta$)，则得到集合 $B=\{Q_{ij}\mid Q_{ij}>\beta\}$。其中 β 为二级阈值，用于发现最可能具有语义关系的关键词对，取值范围为(1>β>α>0)。

Step4：若集合 B≠ϕ，则 NewSemanticSet←B，否则 NewSemanticSet 为空。

4 实验

4.1 实验与分析

为了便于与传统的基于关键字匹配的信息检索和基于概念的信息检索进行对比，本文采用信息检索系统的一般评价标准，利用查准率和查全率来定量分析 QSEBMI 较传统的问句扩展系统的优劣，并用 F 量度(F-measure, Van Rijsbergen, 1979)来综合查准率与查全率指标：

关于查准率、查全率以及 F 量度的计算公式如下：

$$precision=\frac{|A\cap R|}{|A|}\quad recall=\frac{|A\cap R|}{|R|}$$

$$F=2*(recall*\ precision)/(recall+precision)$$

A 表示信息检索系统获取的数据记录的集合；

R 表示数据全集中所有与用户查询相关的数据记录的集合。

本文的实验主要针对特定的语料库进行检索测试。我们用 Spider 从网上获取了计算机类文档集作为实验数据，共计 200 余篇，文章中语言为中文，且长度不等。我们手动准备了 20 余条真实的查询问句，来针对不同的计算机领域进行查询。在权值的计算中参数取值如下：$\alpha=0.4$，$\beta=0.8$。我们分别将本文所提出的 QSEBMI 模型与不进行查询词扩展的方法、局部上下文分析方法进行比较，将其得出的查全率和查全率以及其 F 量度值进行比较，如表 1 所示。

表 1

类型	MI based query semantic expansion			MI and Ontology Based CR Model		
	Precision	Recall	F-measure	Precision	Recall	F-measure
电脑硬件设备	0.303	0.323	0.313	0.402	0.351	0.338
计算机软件	0.293	0.257	0.274	0.330	0.287	0.273
Total averaged	0.298	0.290	0.293	0.311	0.296	0.301

由上述的实验结果可知，QSEBMI 模型通过引入“互信息”与“用户兴趣模型”并使用了改进算法后，较传统的问句扩展系统在查准率、查全率和 F 量度上都得到了一定的改善。

4.2 筛眼现象

从本体的完备度与新语义关键字的关系曲线我们可以看出(如图 3 所示)，过滤之后得到的互信息矩阵与领域本体的完备性有密切的关系。当领域本体的完备性不高时，所得到

(下转第 177 页)

信息系统的分配协调集以及分配约简的概念。当(U,AT,D)是不完备单模糊目标信息系统并且$\alpha\geqslant\beta$时，(α,β)分配协调集便是文[15]中提出的(α,β)协调集，故此时的(α,β)分配约简为文[15]当中的(α,β)精度约简；当(U,AT,D)是完备的符号值目标信息系统，并且当$\alpha>0$时，α水平下分配协调集便是文[18]当中提出的分配协调集，从而α水平下分配约简为文[18]当中提出的分配约简，所以本文提出的分配约简是不完备单模糊目标信息系统的(α,β)精度约简概念的推广，同时也是完备的符号值目标信息系统的分配约简概念的推广。本文还给出了各种分配约简相应的辨识矩阵，从而给出了相应的约简算法，并通过实例表明了算法的有效性。

参 考 文 献

1 Kryszkiewicz M. Rough set approach to incomplete information systems[J]. Information Sciences. 1998, 112: 39～49
2 Kryszkiewicz M. Rules in incomplete information systems[J]. Information Sciences. 1999, 113: 271～292
3 Leung Y, Li Deyu. Maximal consistent block technique for rule acquisition in incomplete information systems[J]. Information Sciences. 2003, 153: 85～106
4 Zhang Wen-Xiu, Mi Ju-Sheng. Incomplete Information System and Its Optimal Selections[J]. Computers and Mathematics with Applications, 2004, 48: 691～698
5 张文修，梁怡，吴伟志. 信息系统与知识发现[M]. 北京：科学出版社，2003
6 Zhang Mei, Wu Wei-Zhi. Knowledge Reductions in Information Systems with Fuzzy Decisions[J]. Journal of Engineering Mathematics. 2003, 20(2):53～58
7 张梅，朱朝晖. Fuzzy目标信息系统的知识发现[J]. 模糊系统与数学，2005，19(1)：121～125
8 袁修久，张文修. 模糊目标信息系统的属性约简[J]. 系统工程理论与实践，2004，5：116～120
9 管涛，冯博琴. 模糊目标信息系统上的知识约简方法[J]. 软件学报，2004，15(10)：1470～1478
10 袁修久，何华灿. 优势关系下模糊目标信息系统约简的辨识矩阵[J]. 空军工程大学学报(自然科学版)，2006，7(2)：81～84
11 陈德刚，刘民，吴澄，李法朝. 模糊信息系统的代数结构及其约简[J]. 清华大学学报(自然科学版)，2003，43(9)：1233～1235，1264
12 管涛，薛亮，冯博琴. 模糊信息系统上的粗糙约简[J]. 西安交通大学学报. 2005，39(6)：574～577，632
13 Wei Dakuan, Zhou Xianzhong, Dongjun Xin, Zhiwei Chen. Variable Rough Set Model and Its Knowledge Reduction for Incomplete and Fuzzy Decision Information Systems[J]. International Journal of Information Technology, 2006, 3(2): 140～144
14 Wei Da-kuan, Zhou Xian-zhong. Rough Set Model in Incomplete and Fuzzy Decision Information System Based on Improved-Tolerance Relation[C]. In: 2005 IEEE International Conference on Granular Computing, Tsinghua University, China, July 2005. 278～283
15 魏大宽，黄兵，周献中. 不完备模糊目标信息系统粗集模型与知识约简[J]. 计算机工程，2006，32(8)：48～51
16 魏大宽. 基于相似关系的不完备模糊决策信息系统知识约简[J]. 湖南师范大学自然科学学报，2006，29(2)：18～23
17 魏大宽，周献中，黄兵. 不完备模糊决策信息系统的粗集模型与精度约简[J]. 计算机科学，2006，33(6)：182～185
18 张文修，米据生，吴伟志. 不协调目标信息系统的知识约简[J]. 计算机学报，2003，26(1)：1～7

(上接第 167 页)

的关键字集将不能很完善地表达用户的意图，将会有太多具有意义的关键字被漏掉，因此会产生大量的新语义关键字，从而加重了语义发现模块的负担，影响了系统的效率；反之，则会使语义发现模块失去存在的意义，因此我们将这一问题称为“筛眼现象”。

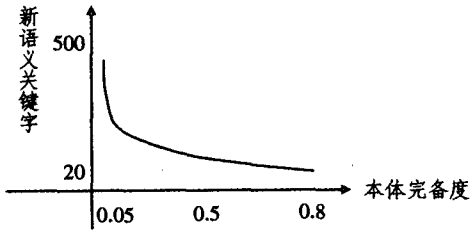

图 3 冗余互信息曲线图

其数学表达如下：设过滤函数 Fi，领域本体 Do，互信息矩阵 Mo，过滤后的互信息矩阵为 Mf 则有：$Mf=Fi(Do,Mo)$。互信息矩阵的过滤完全取决于其完备性，令：W 为 Do 的完备性，取值范围为$[0,1)$，Ri 为 Mf 相对于 Do 而言的冗余互信息量，Roi 为 Mo 相对于 Do 而言的冗余互信息量，Ufi 为 Mf 针对 Do 而言的有用互信息量，Uoi 为 Mo 针对实际语义使用而言的有用互信息量，$Ufi\subseteq Uoi$ 。有以下两条定理：

定理 1 无论 W 取值如何，Ri 趋近于 0。

证明：设过滤函数 Fi，领域本体 Do，互信息矩阵 Mo，过滤后的互信息矩阵为 Mf 则有：

$Mf=Fi(Do,Mo)$。

$\because Mo=\{Uoi\cup Roi$ 且 $Uoi\cap Roi=\phi\}$，Uoi 表示矩阵 Mo 当中的有用互信息量，Roi 表示矩阵 Mo 当中的冗余互信息量。

$\because$领域本体集合 Do 具有准确的语义表述能力。

$\therefore Mf\subset A\subseteq Do$，$Mf$ 相对于 Do 的冗余度 $Ri\to 0$(定理 1 得证)。

定理 2 $Ufi=W*Uoi$

证明：$\because$令领域本体集合 Do 的冗余度为 W，其取值范围为$(0,1)$

$\therefore$当 $W\to 1$ 时，$\lim\dfrac{A-Mf}{A}=0$，即$(Ufi=A-Mf)\to 1$

当 $W\to 0$ 时，$\lim\dfrac{A-Mf}{A}=1$，即$(Ufi=A-Mf)\to 0$

由上两式可得：Ufi 与 Do 的完备性 W 成正比，也就是对互信息矩阵的过滤完全取决于 Do 的完备性和互信息矩阵当中的有用互信息量 Uoi。

因此我们可得以下公式：$Ufi=W*Uoi$ (定理 2 得证)。

结束语 本文提出了一种基于互信息与本体相结合的协同检索模型。该模型将互信息与领域本体相结合，利用互信息与本体在本质上的互补性，从而在查准率、查全率和 F 量度上都优于文[2]中的 QSE_BMI 模型，性能都有所改善。接下来的工作，我们将主要集中在对本体净化与本体补充进行进一步的研究，为实现本体的半自动学习和更新做准备。

参 考 文 献

1 袁占亭，张爱民，张秋余. 基于概念的 Web 信息检索[J]. 计算机科学，2004，31：13～16
2 夏磊，周竹荣. 基于互信息的问句语义扩展研究[J]. 计算机工程与设计，2008. 3
3 王晓珑，关毅，等. 计算机自然语言处理[M]. 清华大学出版社，ISBN：7-302-10089-6，2005. 18～21
4 刘炜，李大玲，等. 基于本体的元数据应用[M]. 上海图书馆
5 Chun-XiaZhang, Cun-GenCao. Domain-Specific Formal Ontology of Archaeology and Its Application in Knowledge Acquisition and Analysis[J]. Computer-Science, 2004(3): 290～301
6 Guarino N. Semantic Matching: Formal Ontological Distinctions for Information Organization, Extraction, and Integration[C]. In: Pazienza MT, ed. Information Extraction: a Multidisciplinary Approach to an Emerging Information Technology, Springer Verlag, 2001. 139～170
7 徐立广，金芝. 一个本体评议及本体构造工具的设计[J]. 计算机工程与应用，2006，25：74～79