基于形式概念的语义网本体的构建与展现*)

徐红升[1] 沈夏炯[1,2] 刘宗田[2]

(河南大学计算机与信息工程学院河南大学数据与知识工程研究所 河南 开封 475001)[1]

(上海大学计算机工程与科学学院 上海 200072)[2]

摘 要 作为语义网基础的本体是共享概念模型的明确的形式化规范说明,它提供一种让计算机可以交换、搜寻和认同文字信息的方式。有效地构建、展现本体成为应用本体的关键问题,然而,现有构建本体的各种方法都在不同方面存在着限制。经过分析比较,本文采用形式概念分析理论构造本体阶层来弥补缺陷,并结合机率模式展现本体,用于表达概念之间及概念、资料间的相关性,利用文件与概念的相关性排序结果,以便于用户找到最相关的信息,从而有效地提高了信息查找的效率。本文通过实例来演示本体的构造与表达。

关键词 本体,概念格,形式概念分析,相关性

Construction and Presentation of Ontology on Semantic Web Based on Formal Concept

XU Hong-Sheng[1] SHEN Xia-Jiong[1,2] LIU Zong-Tian[2]

(School of Computer and Information Engineering, Henan University, Institute of Data and Knowledge Engineering, Henan University, Kaifeng, Henan 475001)[1] (School of Computer Engineering and Science, Shanghai University, Shanghai 200072)[2]

Abstract As the foundation of the semantic web, ontology is a formal, explicit specification of a shared conceptual model and provides a way for computers to exchange, search and identify characteristics. It becomes an important problem on ontology application to build and present ontology. However, there are some defects between several kinds of existing ontology construction methods in many aspects. After analyzing and comparing, this paper makes up these defects by applying formal concept analysis theory to construct concept hierarchies of ontology, and expresses the relevancy of concepts and documents in combination with probabilistic model for ontology presentation. Users can easily search for the most relevant information using the order of the related documents and concepts, so that search efficiency is improved. A case is given to demonstrate how to construct and present ontologies.

Keywords Ontology, Formal concept analysis, Concept lattice, Relevancy

1 引言

目前,因特网在信息表达和检索方面存在着缺陷,比如查全率、查准率、检索速度及客户响应时间尚不能很好地满足用户的需要,主要原因在于其设计目的是面向用户直接阅读与处理,而没有提供计算机可读的语义信息,因此限制了计算机在信息检索中的自动分析处理以及进一步智能化的信息处理能力。解决这些问题的方法就是语义网[1]的构想,目的是为因特网上的信息提供具有计算机可以理解的语义。语义网将本体建构在 XML 和 RDF 等网络语言上,使得本体不再只是表达词汇之间的关系,还包括词汇和资源之间的关系(资料的目的、描述资料的信息)以及资源和资源之间的关系。因而本体的目标更加明确和详细[2]:用于获取、描述和表达相关领域的知识,提供对该领域知识的共同理解,确定该领域内共同认可的词汇,并从不同层次的形式化模式上给出这些词汇和词汇间相互关系的明确定义。

本体的构造问题已受到人们的广泛重视,语义网上本体的建构,都是在网站建设初期就建立好网站信息所代表的本体,需要依赖领域专家定义好类别架构。一般认为人工定义类别架构会存在以下问题:①类别架构不够灵活。无论资料库形式或是目录形式的网站,都是根据事先定义好的类别架构对信息做分类工作,然而随着信息的快速增加与变动,就会出现人工定义的类别架构不能适应信息内容的变化而自动产生出相应的类别;②现有的类别架构所表现的类别相关性通常是绝对的,无法表示出相同类别层次级的相对重要性,导致概念搜索的路径固定且容易遗漏重要的信息。现有的本体建构技术,如字典、文字分类、知识库等构建方法都不能解决这些问题。而形式概念分析[3]是一种从数据集合中发现概念结构的形式化分析方法,用于概念的构造,排序和显示。所生成的概念格作为数据分析与知识发现的有效工具,具有明确的层次关系,含有丰富的语义信息。同时概念格 Hasse 图能清晰地表达概念与概念之间的层次结构,即本体的阶层,本文采用形式概念分析方法来构造本体阶层,并以机率模式来表达概念间的强弱关系及概念与资源的相关性。

2 形式概念分析

形式概念分析(Formal Concept Analysis)是应用数学的一个分支,它来源于哲学领域对概念的理解。根据用二元关系来表达领域中的形式背景,从中提取概念层次结构,即概念格,从数据集中生成概念格的过程实际上是一种概念聚类的过程,它的每个节点被称为一个概念,概念被理解为由外延和内涵两部分组成。概念的外延指属于这个概念所有对象的集合,内涵则是指所有这些对象所共有的特征(或属性)集。目前概念格理论已被应用于机器学习,数据挖掘,信息检索等方面。

定义 2.1 一个形式背景(formal context)是一个三元组

*)本文由国家自然科学基金项目(60575035)支持。**徐红升** 硕士研究生,主要研究领域为软件工程、网络应用和知识发现;**沈夏炯** 副教授,研究方向是软件工程、知识发现、分布式/并行计算及分布式存储;**刘宗田** 博士生导师,教授,主要研究领域为人工智能和软件工程。

$C=(A,B,R)$，其中 A 是对象的集合，B 是属性的集合，R 是 A 和 B 之间的一个二元关系，即 $R\subseteq A\times B$。aRb 表示 $a\in A$ 与 $b\in B$ 之间存在关系 R，读作对象 a 具有属性 b。

在形式背景中，可以定义两个映射 f 和 g 如下：

$\forall A_1\subseteq A: f(A_1)=\{y\in B\mid \forall x\in A_1, xRy\}$

$\forall B_1\subseteq B: g(B_1)=\{y\in A\mid \forall x\in B_1, xRy\}$

这两个映射被称为 A 的幂集和 B 的幂集之间的 Galois 连接。二元组 $(A_1,B_1)\in P(A)\times P(B)$，如果满足 $A_1=g(B_1)$ 及 $B_1=f(A_1)$，则被称为形式背景 C 的一个形式概念，A_1 称为外延，B_1 称为内涵，C 的所有形式概念的集合记为 $F(C)$。

对于形式概念 (A_1,B_1)、$(A_2,B_2)\in F(C)$，如果 $A_1\subseteq A_2$ 或者 $B_2\subseteq B_1$，则称 (A_1,B_1) 是 (A_2,B_2) 的子概念，称 (A_2,B_2) 是 (A_1,B_1) 的超概念，记为 $(A_1,B_1)\leqslant(A_2,B_2)$。偏序集 $(F(C),\leqslant)$ 构成一个格，称为 C 的概念格，记为 $L(C)$。概念格及其相应的 *Hasse* 图形式，反映了一种清晰的概念层次结构，本质上体现了实体（对象、文件、记录）和属性（特征、词汇）之间的联系，概念内涵和外延的统一。

目前，比较传统的造格算法可以分为两大类：批处理算法和渐进式构造算法。本文主要采用基于渐进式的 Godin 算法[4]来构造概念格，其概念聚类的基本思想是：将当前要插入对象的内涵和格中所有结点内涵进行交集运算，根据交集的结果执行不同的操作。经过建造的概念格可以图形化出相应的 Hasse 图，使概念格模型具有更直观地表达概念层次关系的可视化效果。

3 本体建构技术分析

3.1 现有的建构方法

目前，本体的构建技术还处于发展、完善阶段，并不存在一种广泛应用的方法。但是 Maedche 和 Staab 研究表明[5,6]：建构本体可以借鉴其它领域如语言学、信息萃取、机器学习、数据挖掘及软件工程中的相关技术及研究，并提出了将本体学习分为四个部分：萃取、修正、精练及重复利用。其中萃取部分作为建构本体的核心，成为本文关注的重点，经过研究发现，目前本体的建构方法主要为三大类型，分别简单介绍如下。

3.1.1 基于字典的建构方法

鉴于传统字典对于每个字词所定义的同义词、字根、原形等关系，该建构方法就是利用这种词汇与词汇之间的关系—上位词、下位词来确定概念的阶层关系[7,8]。基于字典的建构方法是其它建构方法的基础，然而以此方法建构的本体通常为一般性的描述，并不是与特定领域相关的本体，因此必须结合其它方法以及由领域专家的参与才能形成有意义的本体架构，故此方法无法独立使用；另外，该建构方法不仅受限于字典本身的范围大小，而形成不同范围的子领域，还存在无法适应环境变化的要求而造成遗漏信息。

3.1.2 基于文字分类的建构方法

以文字分类来建构本体概念阶层的方法是：将相关的词依照同义词间存在的相似概念而类聚在一起，在每类之中有较高频率的词将被选出代表这个类，如此重复地筛选后，最终得到词汇的阶层。现在文字分类方法仍旧存在着一些问题[9]：虽然文字分类通常被视为一种比较客观的建构方法，但不同的人对同一份文件可能有不同的看法，会产生不同的分类标准，因此很难得出一个明确定义的分类，造成输出结果无法解释。除非结合了特定领域专家的参与，常见的解决方法为根据分类结果产生分类规则，然而这样会出现过多的分类规则及特征，最终这种方法是无效的；此外，对于文件的每个字词都会被视为这个文件的属性来计算，因此文字分类一般需要在高维度空间做分类计算，但是经由实验及数学分析证实在高维度空间中计算分类的处理效率低，主要是由于每一个信息点到其它信息点具有相似距离的趋势。若为了提高效率而降低计算维度空间，则会遗失重要信息。

3.1.3 基于知识库的建构方法

以知识库为基础的建构方法必须首先建构相关领域的知识库，建构成的知识库主要由一些基本规则和简单范例组成。当使用者要寻找相关资料时，首先键入关键词，然后利用知识库内的这些规则过滤掉大量的不相关资料，并且利用相类似的范例做比对，抽选出最相关的结果，最后利用知识库内的规则重新建构相关的本体并且可以给予摘要与结果。这种方法与其它三种方法的区别是：知识库内现有的规则也可以认为是本体的一种呈现，而只是再利用这些知识库内已有的规则去组合相关的小型本体[10]。

3.2 建构方法比较

根据对以上 3 种方法的说明，将本体建构的各种技术缺陷总结如下：以字典为基和以知识库为基的方法都存在建构时需依赖特定领域的资料及无法独立使用；以文字分类为基的方法存在输出结果无法解释，在处理效率低和信息遗失两方面，容易顾此失彼。

而形式概念分析的原理是根据由对象、属性所组成的二元关系来描述形式背景，并从中自动抽取出完整的概念层次结构，形成概念格，它是一种精确的数学模型，主要由前驱-后继关系来形成概念格中的结点关系，从本质上反映了实体-属性联系，其相应的 Hasse 图可以准确图形化出本体中的概念层次结构。因此该技术已经被广泛地应用到知识发现和数据挖掘等领域并收到了比较好的效果。本文考虑运用 FCA 技术来构造本体可以有效地突破现有本体构造方法所带来的问题，使构造的本体可以随着信息的变化而不断更新本体结构，从而更新类别架构，满足现在信息快速增长的特点。

4 关于概念之间相关性的机率模式

经过研究相关文献[11,12]，为了得到概念间的相关性数据，通常要建立一个概念相关性网络模型。一般意义上说其中的概念结点应该是能够代表文件的重要词汇，同时这些词汇具有较高的 $TF*IDF$ 值，TF 与 IDF 的积是词汇对文件关联程度的数值表示，所谓 TF(Term Frequency)代表文件中某词汇出现的次数，IDF 代表在文件集中词汇对于特定文件的特殊性。对于 IDF，首先计算 $DF(w)$ 表示所有存在词汇 w 的文件总数，通过公式：$IDF(w)=\log\frac{N}{DF(w)}+1$ 来计算 IDF，N 指文件总数，$IDF(w)$ 值的高低说明了词汇 w 具有文件区分能力的强弱[12]。经过计算每个词汇的 $TF*IDF$ 值，可以筛选出与文件关联程度高的词汇，通常删除最后百分之十的字词再循环计算，直到所设定的字词数目下限，形成词汇集合。采用本体构造技术形成本体架构，然后根据 Kang 等人[11]所提出的方法来计算本体中概念之间的相关性，从而表达概念间的强弱关系，计算公式如下：

$$f_{jk}=relevancy(T_j,T_k)=\frac{\sum_{i=1}^{n}d_{ijk}}{\sum_{i=1}^{n}d_{ij}}\times WeightingFactor(T_k) \tag{3.1}$$

$$d_{ijk}=tf_{ijk}\times\log_{10}\left(\frac{N}{df_{jk}}\times w_j\right) \tag{3.1.1}$$

$$d_{ij}=tf_{ij}\times\log_{10}\left(\frac{N}{df_j}\times w_j\right) \tag{3.1.2}$$

$$WeightingFactor(T_k)=\frac{\log_{10}\frac{N}{df_k}}{\log_{10}N} \quad (3.1.3)$$

公式(3.1)描述两个词汇之间的关联程度，而且关联程度具有方向性，以不同词汇为中心所计算出来的相关性是不相同的。以公式(3.1)为例，即以 T_j 为中心时，计算 T_k 与 T_j 的关系。公式(3.1)可拆分为(3.1.1)、(3.1.2)及(3.1.3)3个子公式。这3个公式也都利用了 TF-IDF[13] 理论。对公式中的变量简单说明如下：n 代表词汇的总个数，N 表示文件的总数，w_j 表示不存在词汇 j 的文件次数。在公式(3.1.1)中，d_{ijk} 由词 T_k 及 T_j 所同时出现的次数及不存在某文件中的文件次数来决定，tf_{ijk} 表示词汇 j 与 k 共同出现在文件 i 的次数，df_{jk} 则表示词汇 j 与 k 同时出现的文件总数；在公式(3.1.2)中，d_{ij} 由词汇 j 对于文件 i 的相关性决定，tf_{ij} 即文件 i 中词汇 j 出现的次数，df_j 即所有存在词汇 j 的文件总数。当两个词汇具有较高的关联度时，在同一份文件中它们出现的次数也应该很高且集中在某些特定的文件之中。此外，在公式(3.1.3)中，$WeightingFactor(T_k)$ 表示词 T_k 对文件的特殊性，词汇 T_k 越普遍时，则 $WeightingFactor(T_k)$ 的值就会越小。

5 本体的构建与展现

在语义网中构建本体，首先要对不同来源的文件资料依次进行去除文件格式、语汇分析、去除 Stop word、去除衍生词、同义词典等适当的处理，以取得最终的词库(Lexical DB)集合。这五种形式的处理是用来筛选出能够代表文件资料的字词集合，以形成文件和词汇的二元关系表。

有关这些文字处理，参考 Ricardo 和 Berthier 提出的方法[14,15]：去除文件格式就是去除与本文不相关的信息，如排版格式、批注等额外信息，形成以纯文字(字符)组成的资料流(Data stream)；语汇分析是将由字符所组成的资料流转换成为字词(Term)集合。英文的语汇分析利用空白或标点符号将资料流转换为字词集合，中文的语汇分析则由中文词知识库工具来转换；在语汇分析中，可能出现很多的字词没有实在意义，通常都是冠词、介词及连接词等与文件概念无关的字词，如 a、an、as、and 等。通常在一份文件中，有超过 80% 的字词是不具有任何意义的，在分析过程中应该被过滤掉。另外，由于原型词的基本意义不变，可以用词的原型来取代它不同的形态，例如 connect，它的变形可以为 connecting、connection 及 connections 等，这样可以减少存储空间及降低计算的复杂度；最后，不同的字词可能代表相同的意义，如宿舍与寝室，因此利用同义词典将多余的同义词去除。

在上述处理的基础上，配合利用 TF-IDF 计算字词的重要性以选择适合的字词集合，结合相应的文件集合形成词汇、文件的二元关系表。下面通过一个实例来说明如何构建本体阶层构架，概念间的强弱关系如何表达及使用者如何通过相关性 Hasse 图快速寻找到想要的资料。

综上所述，FCA 是一种优良的形式化分析方法，且配合渐进式造格算法可以构造出概念格，本文采用该技术来建立概念与概念之间的阶层关系。这里假设已经筛选出能代表文件的词汇集合，将文件集合与词汇集合做对应，若某文件内含有某词汇，则标注上该文件包含的词汇个数，如此可以产生带属性值的二元关系矩阵如表1。本体的形式背景(Context)定义为 L，本体的相关文件集合定义为 G，本体的相关词汇集合定义为 M，文件与词汇的相互关系则为 I，因此以上的关系式可以表示为 $L:=(G,M,I)$，$G=\{1,2,3,4,5,6,7,8,9,10\}$，$M=\{A,B,C,D,E,F,G,H,I,J,K,L,M\}$。

表1 词汇与文件的二元关系表

文件	词汇												
	A	B	C	D	E	F	G	H	I	J	K	L	M
1		4		6							5	4	
2	7			9					14		9		
3		1	5							6		3	
4		2	7							9		5	
5				3	2						5		7
6		7		7							3	6	
7			8		4						4	7	
8			3				2				5	6	
9			3			3				8			4
10			8				2			10		4	

从上面由文件、词汇构成的二元关系表来看，词汇属性都是用具体数值表示的，不是简单的选择标记。因此表1是多值形式背景，需要转换成单值形式背景才能用 Godin 算法造格，表1中用数值表示属性的含义是：某文件具有某词汇属性并且用数值来标示所含有的词汇个数，所以在进行转换时，只需将表中的数值用选择标记来代替，如："√"，这样就可以将表1转换成单值形式背景，而在以后进行概念间相关性计算时，才关注表1中的具体数值。限于篇幅，这里省略了转换后的单值形式背景，针对单值背景可以采用 Godin 算法进行造格，得到图1。

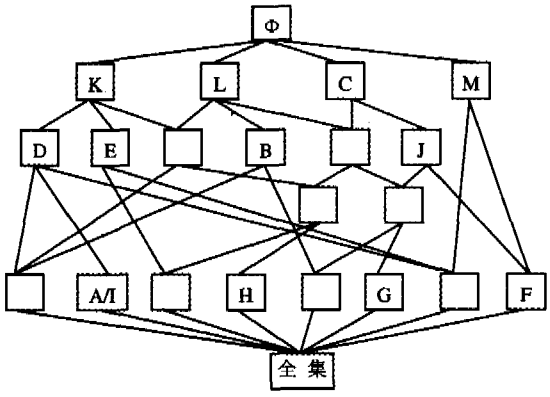


图1 本体的概念图

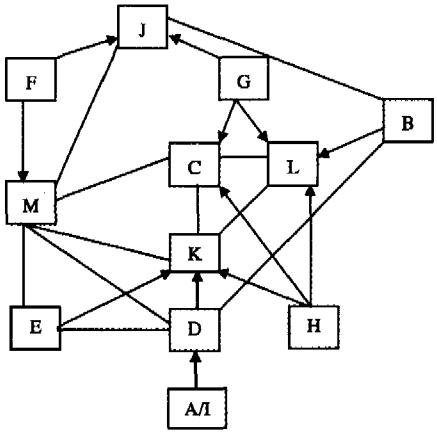


图2 本体的概念关系图

图1采用简化的标注方法，原则为：在图中同一个词汇属性的节点标注只出现一次，最上层的节点是唯一的，其内涵为

空集，而外延则包含了所有对象文件；最下层的节点也是唯一的，其内涵包含了所有词汇属性，而外延则为空集，这样可以更清晰地查看词汇所代表的概念之间的阶层关系。

为了清楚地说明词汇所代表的概念在本体中的相互关系，可以忽略本体概念图中的空概念，直接以 is-a 的关系表示，我们用带箭头的线段来表示两个概念的前驱-后继关系，箭头所指的是前驱概念，箭尾是后继概念。用无方向线段来表示两个概念相互交错关系(a-part-of)，如词汇 D 所代表的概念和词汇 E 所代表的概念没有前驱-后继关系，但文件 5 同时描述这两个概念，故 D 和 E 用无方向线段连接。最后一种是两个概念间没有任何连线，说明这两个概念是独立的，它们可以通过其他的概念发生联系，如图 2。

在生成了概念间的层次关系后，对于非阶层式的关系，即相互交错的概念，我们需要表达概念与概念之间的相对关系。这里采用上述的机率模式来计算概念间的相关性，在图 2 的基础上，通过表 1 加上公式(3.1)、(3.1.1)、(3.1.2)、(3.1.3)来计算概念权重关系，得到图 3。

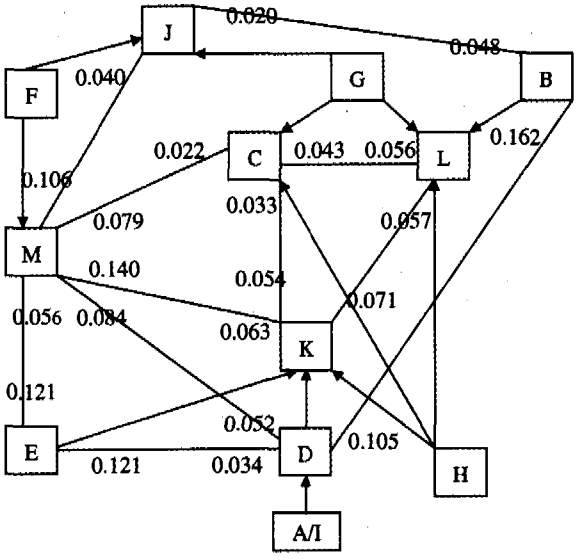
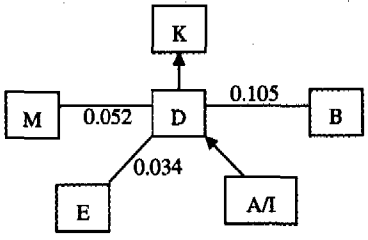


图 3　本体中的概念关系权重图



文件名称	相关度
文件2	0.226
文件6	0.211
文件1	0.198
文件5	0.12

图 4　以概念 D 为中心的概念关系图

图 3 是对概念格相应的 Hasse 图进行简化和忽略空概念等操作后，并结合机率模式所形成的关系模型，图中用具体的权重数据精确地量化了非继承关系概念间的相关程度，即标注在线段上的数字。使构造后的本体概念图更加准确地表达概念间的关系，至此我们产生出本体的完整架构。

最后一步就是展现本体，目的是能够让使用者做相关概念的搜寻及浏览，提供给使用者清晰的概念关系，方便用户查询。如在图 3 中，当使用者选择了某一个节点时，则会呈现出这个节点的上、下阶层概念，相关的概念节点及列出与此节点相关的文件，供使用者参考。以机率模式来表达的图 4 为用户点选了概念 D 后的展现。

在图 4 中，A/I 节点为概念 D 的子概念，即箭尾方向的概念节点，K 节点为概念 D 的父概念，即箭头方向的节点。概念 M、B 及 E 为与 D 概念相关的概念节点，无方向线段中的数字代表与 D 概念的强弱关系，即以 D 来看 B 概念与其最为相关，值为 0.105，其它依次为 M、E。右边是与 D 概念相关的文件列表，其中文件 2 是与 D 概念最相关的文件，其重要性为 0.226，其它相关性排序依次为文件 6、文件 1、文件 5。另外由公式 3.1，我们也可以计算一份文件在单一概念及多个概念间不同的权重，得到相关性的排序，即以文件为中心，计算与不同概念间的权重关系，可以推荐给使用者。此外，使用者可以依据如图 4 的概念关系图重新选择其它的概念，以挖掘其它相关的概念或是同时选择多个概念以求得多个概念间的关系及相对重要的文件。

结束语　通过比较，本文发现传统本体构建方法在各方面存在着缺陷，而形式概念分析能很好地突破这些限制，故成为本体构造的可行方法。本文采取 FCA 方法构造本体，并配合相关性机率模式来达到本体的构造、表达一体化效果，清楚地表达了本体的层次构架和概念间相关性及概念、资源间重要性关系。结合实例阐明了本体从构造到展现的全过程，在一定程度上提高了本体应用在信息检索中的查找效率。对于本体构造的词汇萃取阶段，如果能有效地提取出代表文件的词汇，将直接影响到以后的工作效率，因此这方面的工作是相关研究的基础。目前对构造本体的方法还没有形成统一的性能评估标准，这还是一个需要进一步研究的方向。

参 考 文 献

1　Berners-Lee T, Hendler J, Lassila O. The Semantic Web. Scientific American, 2001(5)

2　宋炜，张铭. 语义网简明教程. 北京：高等教育出版社，2004. 108～131

3　Ganter B, Wille R. Formal Concept Analysis: Mathematical foundations. Berlin: Springer, 1999

4　Godin R, Missaoui R, Alaui H. Incremental concept formation algorithm based on Galois (concept) lattices. Computational Intelligence, 1995, 11(2): 246～267

5　Maedche A, Staab S. Ontology Learning for the Semantic Web. IEEE Intelligent Systems, 2001, 16(2): 72～79

6　Maedche A, Staab S, Hotho A. Ontology-based text clustering. In: Proc. of the IJCAI-2001 Workshop "Text Learning: Beyond Supervision", August, Seattle, USA, 2001

7　Khan L, Luo F. Ontology Construction for Information Selection. In: 14th IEEE Intl Conf. on Tools with Artificial Intelligence (ICTAI'02), Washington DC, 2002. 122～127

8　Tan K-W, Han H, Elmasri R. Web Data Cleansing and Preparation for Ontology Extraction using WordNet. In: 1th Intl Conf. on Web Information Systems Engineering(WISE'00), 2000, 2

9　Williams A B, Tsatsoulis C. An instance-based approach for identifying candidate ontology relations within a multi-agent system. In: Proc. of 14th European Conf. on Artificial Intelligence, Berlin, Germany, 2000

10　Alani H, Kim S, et al. Automatic Ontology-based Knowledge Extraction from Web documents. IEEE Intelligent Systems, 2003, 18(1): 14～21

11　Kang S H, Huh W, Lee S, et al. Automatic Classification of WWW Documents Using a Neural Network. In: Intl. Conf. on Production Research, Bangkok, Thailand, 2000

12　Neto J L, Santo A D. Document clustering and text summarization. In: Proc. of 4th Intl. Conf. on Practical Applications of Knowledge Discovery and Data Mining (PADD-2000), 2000. 41～55

13　Salton G. Introduction to Modern Information Retrieval. McGraw Hill Book Co, 1983

14　Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval. Addison-Wesley. New York: ACM Press, 1999

15　Wu S H, Day M Y, Tsai T H, et al. FAQ-centered Organizational Memory. In: Matta N, Dieng-Kuntz R. Eds. Knowledge Management and Organizational Memories. Kluwer Academic Publishers, 2002