中国科技资源导刊 ISSN 1674-1544 CHINA SCIENCE & TECHNOLOGY RESOURCES REVIEW
2010 年 7 月 第 42 卷第 4 期 55-60 ISSN 1674-1544 Vol.42 No.4 55-60，Jul. 2010

一种基于知网的文档语义模型构建方法

许 琦
（台州职业技术学院机电工程学院，浙江台州 318000）

摘 要：文章提出一种基于语义知识库知网和向量空间模型理论的文档语义模型构建方法，论述知网知识描述方式的特点，提出一种滑动窗口语义消歧算法，利用知网的义原层次体系对文档模型进行语义化处理，根据语境确定语义，将模型特征项转换为关键词的义项，较好地解决了由于自然语言中存在的同义、近义、上下位等语义关系而产生的模型偏差问题。通过计算义项相似度，加权得到文档相似度。实验证明，该方法较好地描述了文档特征，能够达到良好的聚类效果，是切实可行的。

关键词：语义消歧；知网；向量空间模型；相似度计算
中图分类号：TP391 **文献标识码**：A **DOI**：10.3772/j.issn.1674-1544.2010.04.009

A HowNet-Based Semantic Modeling Method of Document

Xu Qi
(Department of Mechanical Engineering, Taizhou Vocational and Technical College, Taizhou 318000)

Abstract: A semantic modeling method of document is put forward. It's based on the semantic repository HowNet and vector space model theory. The characteristics of knowledge describing manner in HowNet are discussed. A slipped window semantic disambiguation algorithm is put forward where the model is semantic handled using the sememe hierarchical system. The model's terms are transformed into meanings according to context and its deviation problems that caused by semantic relations in the language such as synonymous, similar and hypernym-hyponym are solved. The similarity of documents is weighted by calculating the similarity of meanings. Experiment results indicate that the method accurately describes the features of resources which achieves good clustering effects and is feasible.

Keywords: semantic disambiguation, hownet, vector space model, similarity calculation

1 引 言

网络资源 80% 以上是文本资源，如何用统一的数据模型来准确有效地描述网络文本资源 (以下简称文档) 的特征，以便对这些文档进行有效地组织管理，提高资源效用，是一项富有挑战性、长期、复杂的工作。为此，人们提出了 3 种方法：①基于关键词的建模方法 [1-2]；②基于本体的建模方法 [3-5]；③基于语义的建模方法 [6-8]。与①相比，③从词的语义层次上表达文档特征，以概念作为文档的特征项，较好地体现了自然语言中词语之间同义、近义、上下位等关系，弥补了仅以基于关键词的向量空间模型表示文档的缺

作者简介：许琦 (1983-)，男，硕士，讲师，研究方向：企业信息化、网络化制造。
基金项目：浙江省高校优秀青年教师资助计划项目“面向多终端设备的知识信息服务平台研究及应用”；浙江省教育厅科研项目 (Y200909672)；台州职业技术学院校级重点课题 (2010ZD03)。
收稿日期：2010 年 4 月 21 日。

点。与②相比，③只提取那些对文档建模有用的概念，并不需要完全理解全文的语义，符合当前的自然语言处理水平，而且它没有本体知识库的领域性限制，可以建立通用的数据模型。

现有的基于语义的建模方法大都将文档映射到某个语义知识库中，通过不同的方法提取语义特征，以抑制自然语言中存在的同义、近义、上下位等语义关系而产生的模型偏差，如文献[6]采用奇异值分解的方法提取文档特征，文献[7]以 WordNet 为语义知识库，提取主题句向量组作为文档特征。本文主要考虑词语语义对语境的依赖性，根据词语的语用信息和语法信息确定语义信息，以知网为语义知识库，在向量空间模型的基础上，通过语义消歧来提取文档特征。

2　语义知识库知网的特点

知网是以一个汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间关系为基本内容的常识知识库[9]。与《同义词词林》和 WordNet 不同，知网采用多维的知识形式描述语义。

定义 1 义原：义原是最基本的、不易再分割的意义的最小单位。

定义 2 概念：概念是对词语语义的一种描述。一个词语可能有多种概念。

知网通过知识系统描述语言（Knowledge Database Mark-up Language，KDML）描述概念，体现概念与概念之间以及概念所具有的属性之间的相互关系[10]。KDML 所用的“词汇”就是义原。知网共提取并最终确定了 1500 个义原，这些义原可分为以下几个大类：① Event| 事件；② entity| 实体；③ attribute| 属性；④ aValue| 属性值；⑤ quantity| 数量；⑥ qValue| 数量值；⑦ SecondaryFeature| 次要特征；⑧ syntax| 语法；⑨ EventRole| 动态角色；⑩ EventFeatures| 动态属性。对于这些义原，可以把它们归为三组：第一组，包括①到⑦类的义原，称为“基本义原”，用来描述单个概念的语义特征；第二组，包括第⑧类义原，称为“语法义原”，用于描述词语的语法特征，主要是词性（Part of Speech）；第三组，包括第⑨和第⑩类的义原，称为“关系义原”，用于描述概念和概念之间的关系。义原之间存在着复杂的关系，知网描述了上下位关系、同义关系、反义关系、对义关系、属性 – 宿主关系、部件 – 整体关系等 16 种义原关系，其中最主要的是上下位关系。根据义原的上下位关系，所有的“基本义原”组成了一个义原层次体系（图 1）。义原层次体系是一个树状结构，它是知网描述词语语义的基础。

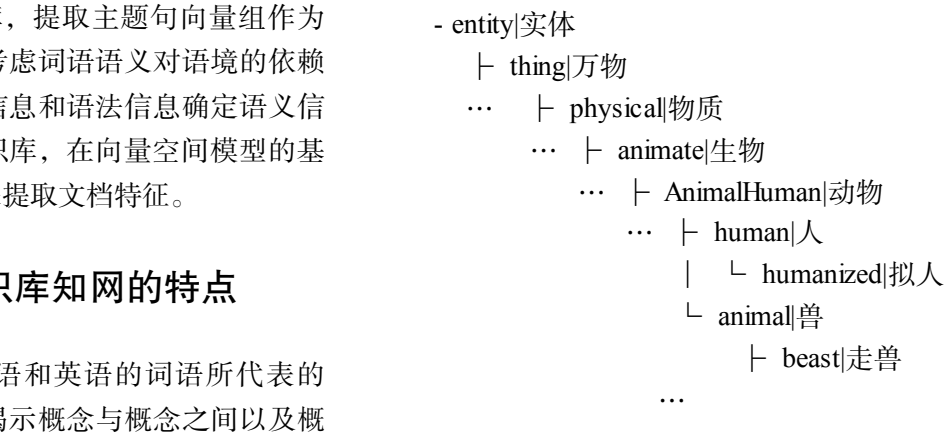


图 1　知网义原层次体系

在《同义词词林》与 WordNet 中，所有词语构成一个树状层次体系，体系中每个结点表示一个概念。一般来说，在一个树状层次体系中，任何两个结点之间有且只有一条路径。于是，这条路径的长度就可以作为这两个概念的语义距离的一种度量。利用《同义词词林》来描述词语的语义，虽然比较准确地反映了词语之间语义的相似性和差异，但前提是树状层次体系是静态的、稳定的。当树状层次体系中的结点发生变化时，则使得路径长度发生变化，进而影响语义之间关系。自然语言是一个动态的知识体系，随着时间、环境的变化而变化，因此基于《同义词词林》或 WordNet 描述词语语义的方法有一定的应用局限性。

知网是以概念和概念之间的关系为描述对象的知识库，与树状的语义知识词库有着本质的不同。知网根据义原的上下位关系构成一个树状的层次体系。每一个概念通过一组义原表示，概念本身并不是义原层次体系中的一个结点，义原才是这个层次体系中的一个结点。义原相对稳定，因此义原层次体系也是稳定的。而且，一个概念

并不是简单的描述为一个义原的集合，而是以义原为基本词汇，通过 KDML 来表达的一个语义表达式。知网“静态地对概念标注，通过概念之间内在的联系动态地、综合地反映它们的关系网络”的这种知识描述方式，较好地解决了语义的静态性和自然语言的动态性之间的矛盾，因此本文以知网为基础来探讨文档语义建模问题。

3　文档语义模型的构建

3.1　模型初始化

采用向量空间模型理论建立文档特征模型，出发点是文档中都包含着能够表达文档内容的独立属性（关键词），文档则表示为这些属性的集合，构成文档特征向量。向量空间模型的概念最早是由 Gerard Salton 提出[11]。他认为，在检索过程中，有必要对信息本身进行分析来帮助检索。这个分析的过程就称为“建立索引”。建立索引主要是为了表现该文档区别于其他文档的内容。假设文档集 D 包含 n 篇文档，表示为 $\{d_1, d_2, \ldots, d_n\}$，建立索引大体上可分为两个步骤：一是经过中文分词、词性标注、词频统计等文档预处理之后，建立特征词向量 $(T_1, T_2, \ldots, T_m)$。在大部分基于向量空间模型的信息检索系统中，特征词一般是指包含足够信息，能反映某一文档 di 内容的关键词。二是赋予各特征词一定的权重（$w_{i1}, w_{i2}, \ldots, w_{im}$），以反映该特征词在区别文档内容上的重要性和价值。文档 d_i 可表示为 $\{T_1: w_{i1}; T_2: w_{i2}; \cdots; T_m: w_{im}\}$，特征项为 $T_j: w_{ij}$，其中 T_j 表示文档 d_i 的第 j 个特征词；w_{ij} 为权重。文档在向量空间模型中以特征向量的形式表示，不仅方便地表现了文档之间的关系，而且更容易计算彼此的相似度。

在向量空间模型中，一个特征词表示空间中的一个维度，文档 d_i 在该维度上的值表示该文档在这个维度上的重要程度，这个值即为权重。权重计算方法一般是将权重分解为两部分分别定义：特征词局部权重和特征词全局权重。特征词局部权重表示特征词对文档的重要性和价值；特征词全局权重则说明特征词在文档集中区别分辨文档的能力是不同的。Gerard Salton 指出了一个经验公式来计算特征词权重[11]，如式（1）所示。

$$w_{ij} = lwt(i,j) \# gwt(j) = \frac{t_{ij}}{\sqrt{\sum_{j=1}^{m} (t_{ij})^2 \# \log^2(\frac{n}{n_j} + 0.01)}} \# \log(\frac{n}{n_j} + 0.01) \quad (1)$$

其中，w_{ij} 表示特征词 T_j 在文档 d_i 中的权重，$lwt(i, j)$ 表示特征词 T_j 在文档 d_i 中的局部权重，$gwt(j)$ 表示特征词 T_j 的全局权重，t_{ij} 表示特征词 T_j 在文档 d_i 中出现的次数，n 为文档总数，m 为特征词总数，n_j 表示出现特征词 T_j 的文档数。该公式主要考虑词频的因素，其基本思想是特征词在某一文档中出现的频率越高而在其他文档中出现的频率越低，则表明该特征词对这一文档的重要性和价值越大，越能反映该文档的内容以区别于其他文档，所以权重越高。

向量空间模型假设特征词之间是线性无关的，而自然语言中存在许多同义词、近义词、上下位词的关系，以具体关键词作为特征词可能引入词汇噪音信息，使得文档模型出现偏差。因此尚需进行语义消歧处理。

3.2　语义消歧

根据全信息理论[12]，自然语言作为认识主体所表述的“事物运动状态及其变化方式”，包括形式、含义和其对认识主体的效用等方面，分别称为事物的语法信息、语义信息和语用信息，而这三者的整体则称为“全信息”。因此不但要关心语言的语法信息，还需要关心语言的语义信息和语用信息。文档是自然语言的载体，同样具有词语同义性、多义性、对短语的依赖性、对上下文的依赖性等特点。根据全信息理论，可以利用词语的语用信息和语法信息，确定词语在文档中所表达的语义信息。

定义 3　语义消歧：根据词的词性及其在文档中的位置获得该词在该文档位置中为某一语义或者倾向于某一语义的过程，称为语义消歧。

基于知网的语义消歧方法如下。

通过词性确定关键词的语义。根据关键词 T 的词性 p，查知网得到该关键词词性为 p 的义项，如果义项个数等于 1，则这个词性为 p 的义项就是该关键词 T 的语义，消歧结束，直接返回；如果义项个数大于 1，则记录关键词 T 词性为 p 的所有义项，并转到下一步。

通过上下文确定词的语义。由于知网中词的语义是由义原定义的，因此基于全信息的概念消歧方法的基本思想就是根据上下文中关键词的义原，对该关键词的义项进行概率统计，与上下文发生关联最多的义项就是该关键词的语义。如果关键词 T 的某个义项的某个义原 S 与上下文中的某个词 Tc 的某个义项的某个义原 Sc 具有以下任意一种关系：相同义原、空间—事件关系、施事 | 经验者 | 关系主体—事件关系、材料—成品关系、事件—角色关系、上下位关系，则认为关键词 T 与词 Tc 有关联，使这两个词发生关联的这两个义原称为关联义原，而这两个义原所在的义项则称为关联义项。使用滑动窗口法确定多义关键词 T 在上下文中词性为 p 的义项，算法描述如下。

算法：滑动窗口法确定多义词 T 词性为 p 的义项

输入：以关键词 T 为中心、窗口大小为 2_{n+1} 的上下文字符串 $T_1\ T_2\ ...T_n\ T_{n+1}T_{n+2}\ ...\ T_{2n+1}$

输出：关键词 T 在上下文中词性为 p 的义项

函数体：

（a）给关键词 T 词性为 p 的各个义项 M_j 赋初始关联度为 0，即 $Relativity(M_j)=0$；

（b）　*for* 窗口中除关键词 T 外的每个词 T_i

（c）　　*for* 词 T_i 词性为 p_i 的每个义项 M_i

（d）　　*for* 义项 M_i 的每个义原 S_i

（e）　　*for* 关键词 T 词性为 p 的每个义项 M_j

（f）　　*for* 义项 M_j 的每个义原 S_j

（g）　　*if*(S_j 和 S_i 关联)

（h）　　$Relativity(M_j)= Relativity(M_j)+1$；

（i）　　*End if*;

（j）　最后，关键词 T 关联度最大的义项 M_j 就是该关键词在上下文中词性为 p 的义项。

实验证明，窗口的大小 (2n+1) 的选取直接影响语义消歧的效果。如果窗口太小，上下文中有关联的词会被排除在窗口之外，不能参与统计；如果窗口太大，窗口内就会有过多没有关联的词，反而增加运算负荷。

语义消歧之后，文档模型的特征词由关键词转换为原关键词在知网中的义项。

至此，文档语义建模过程总结如下所示：

$$\text{文档}d_i \xrightarrow{\text{文档预处理}} \text{关键词序列}T\begin{bmatrix} T_1:p_1:t_{i1} \\ T_2:p_2:t_{i2} \\ \cdots \\ T_m:p_m:t_{im} \end{bmatrix}$$

$$\xrightarrow{\text{计算特征词权重}} \text{特征模型}\begin{bmatrix} T_1:w_{i1} \\ T_2:w_{i2} \\ \cdots \\ T_m:w_{im} \end{bmatrix} \xrightarrow{\text{语义消歧}}$$

$$\text{语义模型}\begin{bmatrix} M_1:w_{i1} \\ M_2:w_{i2} \\ \cdots \\ M_m:w_{im} \end{bmatrix}$$

3.3　相似度计算

词语相似度是指两个词语在不同的上下文中可以互相替换使用而不改变文档的句法语义结构的程度。两个词语，如果在不同的上下文中可以互相替换且不改变文档的句法语义结构的可能性越大，二者的相似度就越高，否则相似度就越低。相似度是一个数值，一般取值范围为 [0,1]。一个词语与其本身的相似度为 1。如果两个词语在任何上下文中都不可替换，那么其相似度为 0。对词语相似度影响最大的是词的语义。

在向量空间模型中，两个文档的相似度采用余弦距离函数来度量。计算公式如下：

$$Sim(d_1,d_2)=\frac{\sum_{j=1}^{m} w_{1j}\times w_{2j}}{\sqrt{\sum_{j=1}^{m} w_{1j}^2}\times\sqrt{\sum_{j=1}^{m} w_{2j}^2}} \qquad (2)$$

该公式基于关键词匹配计算相似度，很大程度上依赖于两向量共有关键词的多少，而没有考虑关键词之间同义、近义、上下位等语义关系，不能很好地反映两文档的语义相关性。

文献 [13] 提出了基于知网的词语语义相似度计算方法，义项 M_1 和 M_2 语义相似度计算公式如下：

$$Sim(M_1,M_2)=\sum_{i=1}^{4}\beta_i\prod_{j=1}^{i}Sim_j(S_1,S_2) \qquad (3)$$

其中，β_i（$1\leqslant i\leqslant 4$）是可调节的参数，$\beta_1+\beta_2+\beta_3+\beta_4=1$，$\beta_1\geqslant\beta_2\geqslant\beta_3\geqslant\beta_4$。$Sim_1(S_1, S_2)$ 表示义

项语义表达式中第一独立义原的相似度，$Sim_2(S_1, S_2)$ 表示义项语义表达式中除第一独立义原以外的所有其他独立义原的相似度，$Sim_3(S_1, S_2)$ 表示义项语义表达式中所有的关系义原的相似度，$Sim_4(S_1, S_2)$ 表示义项语义表达式中所有的符号义原的相似度。义原语义相似度则通过以下公式计算得到：

$$Sim(S_1, S_2) = \frac{\alpha}{d + \alpha} \quad (4)$$

其中 S_1 和 S_2 表示两个义原，d 是 S_1 和 S_2 在义原层次体系中的路径长度，是一个正整数。α 是一个可调节的参数。

笔者认为整体相似度是建立在部分相似度基础上的，即可以通过计算义项相似度，加权得到文档相似度，衡量文档的语义相关性。基于上述考虑，对式（2）作如下改进：

$$Sim(d_1, d_2) = \frac{\sum_{j=1}^{m} w_{1j} \# w_{2j} \# Sim(M_1, M_2)}{\sqrt{\sum_{j=1}^{m} w_{1j}^2} \# \sqrt{\sum_{j=1}^{m} w_{2j}^2}} \quad (5)$$

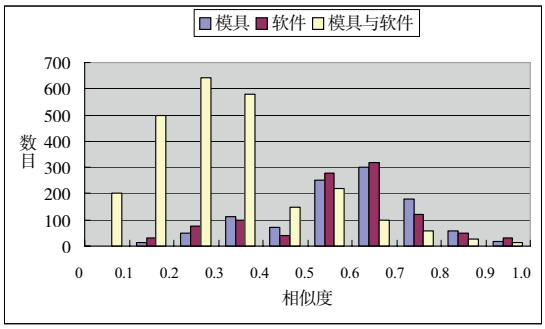
4　实验验证

4.1　实验方法

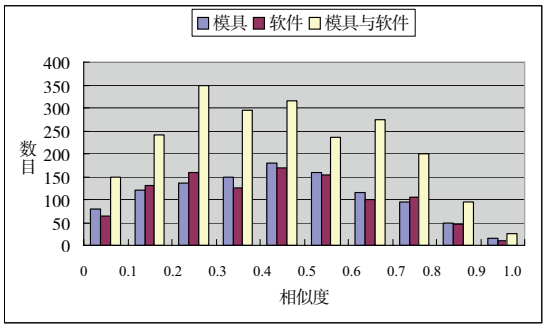
通过以下实验对文档语义建模方法进行验证：以课题组开发的个性化信息服务系统 PISS 为实验平台，以其中的资源收集组件、信息提取组件和资源索引组件等为实验工具，以中国模具网(http://www.molds.cn/default.asp) 为实验数据源，从网站的模具和软件 2 个信息频道中各下载了 50 个样本文档进行实验，将基于文档语义建模方法（以下简称“方法 1”）和基于关键词的建模方法（以下简称“方法 2”）进行比较，以检验两者描述文档特征的能力。

4.2　实验结果

实验样本包括模具和软件主题中各 50 个文档，特征空间总维度为 100。采用方法 1，建立文档语义模型，按照公式（5）分别计算模具主题样本之间 (1225 个计算值)、软件主题样本之间 (1225 个计算值) 以及模具主题样本和软件主题样本之间(2500 个计算值)的相似度，结果如图 2(a)所示。同时，采用方法 2，建立文档关键词模型，按照公式（2）计算模具主题样本之间、软件主题样本之间以及模具主题样本和软件主题样本之间的相似度，结果如图 2(b) 所示。



(a)



(b)

图 2　样本相似度分布

从图中可以看出，采用方法 1 建立的文档语义模型，相同主题样本之间的相似度较大，主要集中在 0.5~0.8 之间；相异主题样本之间的相似度相对较小，主要分布在 0.1~0.4 之间。采用方法 2 建立的文档关键词模型，相同主题样本之间、相异主题样本之间的相似度分布比较均匀，数值差别不大。这表明，方法 1 建立的文档语义模型，较好地描述了文档特征，能够达到良好的聚类效果。

5　结论

用结构化的数据模型来表示网络文本资源，进行有效的组织管理，以实现网络资源效用的最大化，是非常有意义的。本文探讨了网络文本资源的语义建模方法，利用语义知识库知网的语义特性和向量空间模型理论，提出了一种基于知网

的文档语义建模方法。与传统的语义知识库不同，知网采用了 1500 多个义原，通过一种知识描述语言来对每个概念进行描述，这些丰富的语义信息与人的直觉比较符合。文章给出了网络文本资源的语义建模过程，详细论述了建模过程中各阶段的实现方法。通过实验，对方法的可行性和有效性进行了验证。

参考文献

[1] Chen Zhigang, He Pilian, Sun Yueheng, et al. Research and Implementation of Text Classification System Based on VSP[J]. Journal of Chinese Information Processing, 2005, 19(1): 36−41.（in Chinese）
〔陈治刚，何丕廉，孙越恒，等．基于向量空间模型的文本分类系统的研究与实现 [J]．中文信息学报，2005,19(1)：36−41.〕

[2] Liu Shaohui, Dong Mingkai, Zhang Haijun, et al. An Approach of Multi−hierarchy Text Classification Based on Vector Space Model[J]. Journal of Chinese Information Processing, 2002, 16(3): 8−14.（in Chinese）
〔刘少辉，董明楷，张海俊，等．一种基于向量空间模型的多层次文本分类方法 [J]．中文信息学报，2002,16(3)：8−14.〕

[3] Li Jing, Zhou Zhurong, Gan Chengzhi. Research on Ontology Modeling of Learning Resources[J]. Computer Engineering and Design, 2008, 29(1): 251−255.（in Chinese）
〔李静，周竹荣，甘诚智．学习资源的本体建模研究 [J]．计算机工程与设计 , 2008,29(1):251−255.〕

[4] Zhu Jiang, Yu Jie, Zhang Guoning. An Ontology−cloud Based Information Modeling Method of WWW. Computer and Digital Engineering, 2009, 37(9): 12−14.（in Chinese）
〔朱江，俞杰，张国宁 . 万维网基于本体云模型的信息建模方法 [J]．计算机与数字工程 ,2009,37(9):12−14.〕

[5] OuYang Yang，Chen Yufeng，Chen Xiyuan，et al. Ontology Modeling of Domain Knowledge in Semantic Learning Web[J]. Journal of Zhejiang University: Engineering Science, 2009, 43(9): 1591−1596.（in Chinese）
〔欧阳杨，陈宇峰，陈溪源，等．教育语义网中的知识领域本体建模 [J]．浙江大学学报 : 工学版 ,2009, 43(9):1591−1596.〕

[6] Shen Lihong, Zhou Changle. Text Filtering Based on Support Vector Machine of Semantic Space[J]. Computer Applications, 2005, 25(3): 664−665.（in Chinese）
〔沈丽虹，周昌乐．基于语义空间的支持向量机的文本过滤 [J]．计算机应用 , 2005,25(3):664−665.〕

[7] Li Kairong, Lin Ying, Hang Yueqin. Feature Extraction of Document Based on Semantic Model[J]. Computer Engineering and Applications, 2005, 41(17): 173−176.（in Chinese）
〔李开荣，林颖，杭月芹．基于语义模型的文档特征提取 [J]．计算机工程与应用 ,2005,41(17):173−176.〕

[8] Zou Juan, Zhou Jingye, Deng Cheng. A Characteristic Value Extractive Method for Chinese Texts through Semantic Analysis[J]. Computer Engineering and Applications, 2005, 41(36): 164−166.（in Chinese）
〔邹娟，周经野，邓成．一种基于语义分析的中文特征值提取方法 [J]．计算工程与应用 ,2005,41(36):164−166.〕

[9] Dong Zhendong, Dong Qiang. HowNet[EB/OL]. [2009−07−25]. http://www.keenage.com/.（in Chinese）
〔董振东，董强．知网 [EB/OL]．[2009−07−25]. http://www.keenage.com/.〕

[10] Dong Zhendong, Dong Qiang. KDML — Knowledge Database Mark−up Language[EB/OL]. [2009−07−25]. http://www.keenage.com/.（in Chinese）
〔董振东 , 董强 . KDML—知网知识系统描述语言 [EB/OL]．[2009−07−25].http://www.keenage.com/.〕

[11] Salton G，McGill M．Introduction to Modern Information Retrieval[M]．New York：McGraw−Hill Book Company,1983.

[12] Zhong Yixin. Principles of Information Science [M]. 3rd ed. Beijing: Beijing University of Posts and Telecommunications Press, 2002.（in Chinese）
〔钟义信．信息科学原理 [M]. 3 版 . 北京 : 北京邮电大学出版社 ,2002.〕

[13] Liu Qun, Li Sujian. Word Semantic Similarity Computing Method Based on HowNet[C] //Proceedings the 3rd International Conference on Chinese Language Semantics Seminar. Taibei, 2002.（in Chinese）
〔刘群，李素建．基于《知网》的词汇语义相似度计算 [C]// 第三届汉语语义学研讨会论文集．台北 , 2002.〕