HowNet与维基百科知识融合中的义类属性自动构建方法*

崔磊 陈清才† 郭鸿志 王晓龙

哈尔滨工业大学深圳研究生院 计算机科学与技术系 广东 深圳 518055

E-mail: qingcai.chen@gmail.com

摘 要：本文提出了一种开放语义知识库构建方法来融合《知网》和以 Wikipedia 为代表的百科全书，保留《知网》中的语义信息和 Wikipedia 中的丰富资源及其知识框架，通过在两种知识库间建立一个映射关系，构造了一个大规模、带有语义标注的开放语义知识库。知识库以描述类别属性为主，作为知识库构建的重要内容，提出了基于目录词的类别属性提取和约简方法。实验表明，该方法具有较高的准确率，能够准确地提取词条的属性标签信息。

关键词：语义知识库，属性提取，Wikipedia，知网

Auto-Extraction Approach of Semantic Class Attributes in the Fusion of HowNet and Wikipedia

Lei Cui, Qingcai Chen, Hongzhi Guo, Xiaolong Wang

Department of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong 518055

E-mail: qingcai.chen@gmail.com

Abstract: This paper presents an approach for the construction of an open semantic knowledge base based on the fusion of Wikipedia and HowNet. In this knowledge base, each item in Wikipedia is merged into the HowNet semantic category. At the same time, the semantic attributes of each HowNet semantic class are auto-extracted from Wikipedia items that are belonged to the same semantic class. The attributes extraction algorithm based on the catalog words of Wikipedia is proposed for this task. Experiments show that the proposed algorithm reaches high accuracy and is feasible and effective.

Keywords: semantic knowledge base, attributes extraction, Wikipedia, HowNet.

1 前言

语义知识库作为语义分析和计算的基础资源，归纳和描述词语及概念的语法、句法、语义信息和知识表示等，直接影响计算机进行语义理解和分析的准确性。现有的语义知识库建设主要有两种：一种是基于语言学家的人工构造方法，一种是基于特定框架的自动标注方法。前者虽然准确、严谨，但是周期太长，规模很难扩大，后者则面临知识库准确性、合理性验证的难题。近年来，通过庞大的互联网用户群体来志愿参与和构造的网络百科全书 Wikipedia 获得了很大成功，并在短时间内迅速构建了上千万条语言知识条目，通过网友的交互检验(“更多的眼睛发现更多的错误”)，以及在编辑方针上采用中性观点 NPOV(Neutral Point Of View) 等有力措施，这些条目也具有相当高的可信度。显然，在具有数以亿计的

*本文承国家自然科学基金项目（批准号:60703015）和国家 863 项目(项目号: 2006AA01Z197)的资助。

作者简介：

崔磊，硕士研究生，主要研究方向：信息检索，搜索引擎。

† 陈清才，博士，副教授，主要研究方向：自然语言处理、语音处理、信息检索、机器学习，通讯作者。

郭鸿志，博士研究生，主要研究方向：人工智能，信息检索，搜索引擎。

王晓龙，博士，教授，博士生导师，主要研究方向：网络信息处理、人工智能、自然语言处理、生物信息学、声图文智能计算。

用户群体的互联网上，基于志愿者的、开放的、交互式的方法已经成为未来构建大规模、超大规模知识库以及提供准确的信息服务的一种富有成效的模式。这一模式也让我们解决前述规范的大规模语义知识库的快速构建与及时更新问题有了新的途径。

然而，要想将上述新的模式成功应用到语义知识库的建设上，并在此基础上进行语义分析、语义理解，仍然面临诸多挑战。维基百科（Wikipedia）虽然表现形式灵活多样、易于理解，但模式过于宽泛松散，数据非结构化，缺乏语义标注信息，很难直接应用到语义计算中。而与之相比，语义知识库需要更加系统、规范的表达方式，因此也需要更加专业的人员才能参与构建，限制了知识库规模的迅速扩大以及对动态语义变化的快速跟踪。

基于以上问题，本文提出了一种语义知识库构建方法，融合《知网》和以 Wikipedia 为代表的网络百科全书，保留《知网》中的语义信息和 Wikipedia 中的丰富资源及其知识框架，通过在两种知识库间建立一个映射关系，构造一个大规模、具有语义标注的、开放的语义知识库，并着重解决知识库中每个语义类别的属性抽取与扩展问题。虽然知网中的义类本身是具有语义属性的，但是这些语义属性无论对于人们的理解还是知识库的应用来讲都不够具体，我们提出了一种根据归属于每个义类下的 Wikipedia 词条之间的共性来建立更易于理解的义类基本属性的方法。

全文内容组织如下：第二部分简要介绍了国内外语义知识库的研究现状；第三部分讨论了语义知识库的相关概念及其形式化定义；第四部分重点讨论了开放语义知识库的构建过程，给出了词条类别属性提取和约简方法；第五部分对文章提出的开放语义知识库构建方法进行了评测，针对实验结果进行了分析；第六部分进行了总结。

2 相关研究工作概述

语义分析及计算技术的发展带动了相关语义知识库的构建。国内外有代表性的语义知识库有：国外的 WordNet、Frame Net、Mind Net、CYC、Wikipedia 等，中文的《现代汉语述语动词机器词典》、HowNet、CCD、《现代汉语语义词典》、《同义词词林》等。

其中，WordNet[1]主要提供了词语之间同义、同类关系，是基于关系的语义描述；Frame Net[2]利用框架语义学的思想，涉及框架元素、配价、语义关系等；Wikipedia 是一个多语言、动态的全球百科知识全书，采用网络协同创作开发，资源含量巨大；“现代汉语述语动词机器词典” [3][4]将国外语义学理论与汉语的实际情况相结合，涉及格理论；北大 CCD[5]（中文概念辞书）采用类 Word Net 的汉英双语知识描述框架；“现代汉语语义词典”（SKCC）[6]面向汉英的机器翻译,为计算机自动分析、生成汉语、英语句子提供语义信息。

HowNet [7][8]是一个以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的关系为基本内容的知识库，它的一个重要特色就是面向计算机[8]。其语义关系复杂、概念和义元的表示抽象，手工构建的方式导致对新增的词语只能人工地手动添加，很难人工添加修改，无法形成较好的人机互动，缺乏动态性和开放性。

目前的语义知识库绝大多数仍是相互独立的，知识通用性不足，资源重复、合理利用率较低：或语义信息丰富而抽象难懂，导致扩充和修改语义知识库的内容有难度；或规模较小，缺乏通用性；或规模大而语义信息描述较少，不利于计算机的语义分析、理解等。构建一个好的语义知识库要求具有较强的知识互通性、大规模的语料、丰富的语义信息以及较好的人机交互。为达到这样的目标，越来越多的研究人员尝试对多个语义知识库进行有效融合，综合利用各种知识库资源，并在信息抽取[9]、语义信息检索、语义计算[10]等多个领域开展了应用研究。

Suchanek 等基于 Wikipedia 和 Word Net，构造了一个高覆盖率的轻量级的可扩展本体库 YAGO[9]，利用 Word Net 中语义关系从 WikiPedia 中自动抽取事实，用基于规则和启发式的算法构建实体和事实。Torsten Zesch 等人还根据现有的支持英语等四种语言的 Wikipedia 的 API 工具 JWPL 和 JWKTL，抽取分散于 Wikipedia 中的词汇语义信息作为可用的自然语言处理研究资

料[10]。台湾中央研究院研发的中英双语知识本体词网 （Sinica Bow）[11]，以英文 WordNet 和 SUMO（Suggested Upper Merged Ontology）为基础，提供中英文之间的语义转换，语言信息和知识本体的连结，词义的区分、语义关系的描述和多重词义和词义关系的检索，对领域词汇语义知识库的建立，以及人工或自动翻译、自动推理等方面的应用研究起到了很好的推动作用。

3 语义知识库定义

优秀的语义知识库，应该信息涵盖全面，层次分明。本文提出的语义知识库，类别作为主要的描述对象，具有基本属性和语义属性。条目是基本单位，其存在和描述以类别为依存并可具有符合其含义的多个类别。

例如，在语义知识库中存在类别“体育馆”、“巢穴”和词条“鸟巢”，其中词条“鸟巢”既可描述北京奥运会运动馆，又可描述动物的巢穴，按照其含义不同，词条内容包括了以上两个类别的类别信息及属性值，如表 1 所示：

表 1 词条“鸟巢”在不同语义类别下所对应的类别属性

词条	类别	语义属性	类别属性				
鸟巢	体育馆	facilities|设施, @exercise|锻练, sport|体育	建成时间	施工承包单位	坐落位置	设施应用	…
			2006 年建成…	北京城建集团承建 2/3 的工程；…	北京奥林匹克公园内…	北京奥运会运场馆	…
	巢穴	house|房屋, #animal|兽, #alive|活着	位置	形状	类别	特点	…
			…	…	…	…	…

上述义类以及开放语义知识条目可用向量空间的形式进行如下描述[12]：

定义 1：语义知识库词典集 SD 是一个由词条组成的集合，$SD=\{t \mid t=w,c_1,c_2,...,c_k,...,c_j,A_1,A_2,...,A_k,...,A_j,S_1,S_2,...,S_k,...,S_j,w\in W,c_k\in C,A_k\in A,S_k\in S\}$，其中有限集合 W 为概念集；有限集合 C 为类别集；A 为类别的基本属性集的集合； S 为类别的语义属性集的集合；向量 t 为具体的每一词条，t 的分量 w 为词集合 W 的具体取值[12]。

在上述定义的基础上，针对语义知识库的结构特点，我们进一步给出其形式化定义：

定义 2：类别 c_a 是一个由当前类别及其父类组成的一个集合，这里考虑到在开放知识库中允许每个类别（或者条目）存在多个父类，从而我们有：$c_a=\{(c_{a-n},c_{a-p}) \mid c_{a-n},c_{a-p}\in C\}$，其中有限集合 c_a 为第 a 类别集合， c_{a-n} 是第 a 类别及其类别名称， c_{a-p} 为类别 a 的父类别；

定义 3：类别 c_k 的基本属性集合 A_k 定义为：$A_k=\bigcup_{j=1}^{N} a_{kj}$，其中 a_{kj} 为 c_k 类别的第 j 个基本属性，这是开放语义知识库中每个类别下的词条所应该具有的、对语义计算有一定帮助的任何形式的公共性质，与 HowNet 中所定义的义元具有较大区别，该公共属性的抽取也是后文讨论重点。

定义 4： c_k 类别的语义属性集合 S_k 定义为：$S_k=\bigcup_{j=1}^{M} s_{kj}$，这里 s_{kj} 表示在 HowNet 中 c_k 类别的第 j 个义元。

根据上述定义，要构造一个语义知识库框架，就是建立一个从知网、百科到语义知识库的映射关系：$f:C^p\times W\rightarrow C\times(W\cup C\cup A\cup S)$，其中，$f$ 为一个单射，满足百科资源中的每个词条唯一地对应语义知识库中的一个条目； C^p 为原词条的类别，在语义知识库中，该类别被赋予了类别属性，得到新的类别 C；W 为原条目的内容，在语义知识库中，条目与类别及类别属性建立了联系，得到新的条目 $W\cup C\cup A\cup S$。

4 语义知识库构建

4.1 构建语义知识库框架

(1) 百科词典平台

百科词典的基本架构设计，是构建百科平台的第一步。考虑到 Hudong wiki 具有多模块化、功能全面，系统程序设计清晰、便于开发和改进，界面友好、更具人性化，提供在线支持等特点，采用了基于 Hudong wiki 开源知识框架。分类体系，作为知识库的一个重要组成部分，需要满足组织结构清晰、涵盖全面、易于扩充等条件，对比目前流行的开放式百科全书以及语义词典，论文选取 Wikipedia 分类体系作为基础，并进行针对性扩充修改。

(2) 知网与百科词典融合

为使知识库中的条目带有一定的语义信息，论文使用知网对百科词典添加语义描述。在知网（2004版）中，每个词语的概念及其描述形成一个记录。每一个记录包含4项内容，每一项由两部分组成，中间以“=”分隔。每一个“=”的左侧是数据的域名，右侧是数据的值。排列如下：*W_X*= 词语； *E_X*= 词语例子；*G_X*= 词语词性；*DEF*= 概念定义。利用每个记录的“*W_X*= 词语”来标注百科词典的类别，即用知网中的词语匹配百科词典中的类别，为类别赋予语义信息。匹配过程采用字符串模糊匹配算法，这里不做详细阐述。

4.2 类别属性提取及约简

Wikipedia 中的词条，内容丰富，但信息组织松散，遍布到全文当中，这样非结构化的数据十分不利用计算机进行语义分析、检索等。就语义分析、计算而言，Wikipedia 这样内容组织自由分散的形式极需改变，然而大多数词条都是经过多人协作人工编辑的，对绝大多数的普通用户进行内容规范和统一显然是不现实的。

在本文构建的语义知识库中，允许一个词条具有多个类别，词条的内容的描述依赖于类别。类别可将其语义属性和基本属性传递给类别下的词条。一旦词条类别确定，则词条的内容相应确定。因此，类别的基本属性的准确程度、概括程度直接影响整个语义知识库的内容架构，所以提取类别的基本属性十分重要，是语义知识库构建的核心。

传统的文本特征抽取需要综合考虑频度、集中度、分散度 3 项指标[13] ，鉴于当前信息抽取技术的发展情况，直接从大量文本中抽取词条的基本信息构成类别属性，仍旧存在相当大的难度。本文根据百科资源中词条内容的构成特点，即多数词条存在着目录结构及标签词，提出了一种基于目录式的类别属性抽取和约简算法，提取能够准确诠释类别性质的属性集合。针对目录词在一个文本的目录中仅出现一次，算法不考虑目录词的集中度和频度，只考虑其分散度。

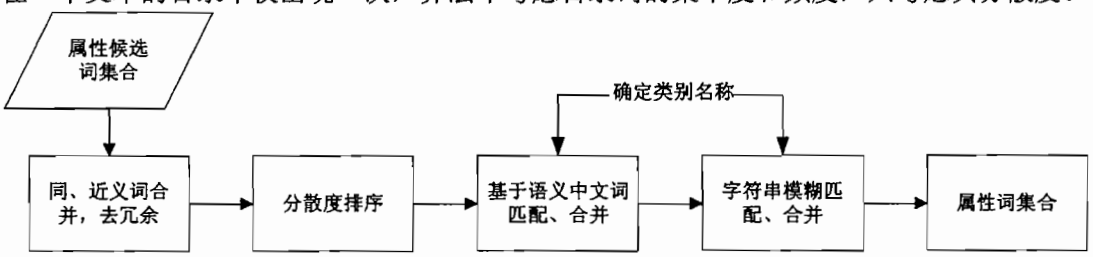

图 1 属性提取及约简

如图 1 所示，属性抽取及约简算法如下：

输入：数据集预处理后得到属性候选词集合；

输出：属性集合

①. 利用《HIT－IRLab 同义词词林（扩展版）》，对属性候选词集合进行同、近义词合并；

②. 按分散度排序候选属性词；

③. 基于语义的中文短语模糊匹配，即用《知网》计算词语相似度：

任意两词语 w_1 和 w_2，在知网中可能存在多条语义信息不同的概念 $c_{11},c_{12},...,c_{1i}$ 和 $c_{21},c_{22},...,c_{2i}$，每个概念可能存在多个义元 S。词语相似度为各概念之间相似度的最大值[14]：

$$Sim(w_1,w_2)=\max_{i=1\cdots n,j=1\cdots m} sim(c_{1i},c_{2j}) \tag{1}$$

其中 $Sim\ (c_{1i},c_{2j})$ 为 w_1 的第 i 个概念和 w_2 的第 j 个概念的相似度。

概念相似度为各义元 S 相似度加权求和[14]，计算公式如下：

$$Sim(c_1,c_2)=\sum_{i=1}^{4}\beta_i\prod_{j=1}^{i}Sim_j(s_1,s_2) \tag{2}$$

其中 s_1 和 s_2 分别是 c_1 和 c_2 的独立义元，$Sim_j(s_1,s_2)$ 为第 j 独立义元的相似度，可由《知网》直接计算得到。$\beta_i(1\le i\le 4)$ 是可调节参数，用以限定各部分重要程度，满足 $\beta_1+\beta_2+\beta_3+\beta_4=1$，$\beta_1\ge\beta_2\ge\beta_3\ge\beta$；

对知网中的未登录词，依照中文短语“重心后移”的特点[15]，对最后位置的词语，依照公式(1)和公式(2)计算相似度。为方便计算，前面的限定性词语不纳入计算；如果相似度大于阈值，则属性合并，分散度加和；

④. 字符串模糊匹配；利用 LCS 对余下目录词求字符串间最大公共子串并计算相似度[16]：

$$Sim\ (w_1,w_2)=|LCS|\Big/(Len(w_1)+Len(w_2)-|LCS|) \tag{3}$$

其中 LCS 为 w_1 和 w_2 的最大公共子串，$|LCS|$ 为 LCS 的长度，$Len(w_1)$ 和 $Len(w_2)$ 分别 w_1 和 w_2 的长度；相似度大于阈值，则属性合并，分散度加和；

⑤. 步骤③、④属性约简的同时，确定类别名称，以分散度最大的目录词命名；

⑥. 每个类别提取前 10-15 个作为基本属性，父类别的属性为其各子类别属性交集；

5 实验与结果分析

5.1 数据集

目前为止，开放语义知识库已收录 Wikipedia 的词条近 20 万，类别数量 1085。为方便测试语义知识库属性自动抽取的准确性，我们暂不考虑各百科词典的分类体系不统一的因素，从语义知识库中选取与其它百科词典公共的 9 个类别作为实验的目标类别。为得到更加全面、知识覆盖率更广的数据集，使用语义知识库中现有条目目录词的同时，爬虫另外从 Hudong wiki、Baidu Baike 中抓取这 9 个类别的文档，合并形成数据集，共 746 篇文档。表 2 描述了数据集中各类别词条及其目录词的分布情况。

表 2：数据集各类别词条及其目录词分布情况

类别	字符集	台风	演员	朝代	城市	花卉	湘菜	犬	运动器材	汇总
词条总数	11	54	159	112	124	112	118	39	24	76
目录词量	11K	49K	159K	124K	221K	187K	106	40K	13K	1450K

数据集获取及预处理过程如下：①. 网络爬虫抓取各类别的词条，得到原始词条集；②.

网页去重、净化，提取标签及目录词得到原始目录词集合；③. 分词、过滤，去除停顿词和体现文本特征的特殊词；④. 返回得到的属性候选词集合。测试的目的是验证生成的属性词是否准确、具有概括性，验证基于目录词及标签提取类别属性算法的有效性。

5.2 评测及分析

由于属性抽取使用语料、提取的方法不尽相同，没有统一的比较基准，本文对语义知识库的评测采用了相对开放的测试方法，对抽取属性的准确率进行评测。考虑到召回率受目录词冗余及表达差异影响较大，缺乏考察价值，对其不做评估。准确率 P(Precision)计算公式如下：

$$P_i = C(c_i)/N(c_i) \tag{4}$$

其中 P_i 为类别 c_i 的准确率，$C(c_i)$ 为正确提取的属性值个数，$N(c_i)$ 为属性值总数；

(1) 数据集全集测试

在得到的类别属性词集合中，属性词比较多，实验表明再次进行属性词过滤及合并，反而会降低属性提取的效果，因此，只选取前 10-15 个属性词作为最终的类别基本属性。计算各类别的准确率，结果如表 3 所示：

表 3 各类别抽取属性准确率

	字符集	台风	演员	朝代	城市	花卉	湘菜	犬	运动器材	平均
准确率（%）	91.67	78.57	91.67	92.31	100	100	100	68.42	100	88.93

由表 3 可知，属性提取及约简算法的准确率比较高，平均达到了 88.93%，在选取的 9 个类别中有 7 类均在 90%以上，其中 4 类的效果达到 100%。由此说明，目录词都是经过多个用户总结概括得到，是人工分析而得到的结果。另外，类别间的准确率依据其数据集语料的具体情况而各有不同。

(2) 增量数据集测试

为考察语料库规模对属性提取的影响，从小到大递增语料数量，提取类别属性。选取类别“台风”、“花卉”、“湘菜”、“朝代”、“犬”，将相应语料分成四等份，每次增加 1/4 进行测试。图 2 显示了增量数据集测试的 5 个类别的属性提取结果：

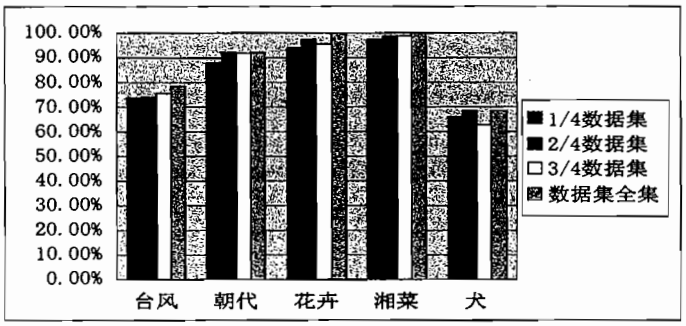

图 2 增量数据集准确率

由图 2 可知，随着语料库规模的增大，基本属性提取的准确率基本均有小幅提高。这说明从统计的角度来看，语料库规模增加的同时，类别下条目的目录词涵盖量增大，属性词所占比例增大，类别的属性描述趋于具体化。因此语料库规模在一定程度上影响基本属性提取和简约方法的准确率。

6 结语

本文构建了一个开放语义知识库，整个知识库以属性描述为主，每个条目以描述类别的基本属性为中心并带有语义属性，添加新条目的同时，利用属性标注词条内容，使词条具有相应的语义标签，主题表达更加鲜明，格式规范，利于词条内容的定向查找和针对某个具体属性的动态修改。利用语义知识库中的《知网》信息可进行词汇的语义相似度计算；语义知识库中词条的类别基本属性结合模糊集等理论可进行文本分类；语义知识库利用其动态、实时性可与信息检索系统形成一定的互动与结合，能够改善搜索引擎的查询性能，得到更加精准的结果。

然而，开放语义知识库仍旧存在着不足，亟待改进。类别属性抽取采用的数据集还存在着目录词冗余等不足，算法的抽取、约简规则也需进一步改进；如何判定类别基本属性的属性值的性质，以及如何与搜索引擎进行结合，通过语义知识库与大规模网页信息的交叉与互动分析，实现基于大规模网页库的知识验证和语义动态特性分析方法，从而建立起一个大规模的、具有快速更新与动态适应能力、语义标注较为准确的语义知识库，都将是下一步研究的重点和难点。

参 考 文 献

[1] Fellbaum. Christiane. Word Net: An Electronic Lexical Database, MIT Press. 1998

[2] Baker, Collin F., et al. The Berkeley Frame Net Project, In Coling'98. 1998. Pages:86-90.

[3] 陈群秀等. 现代汉语述语动词机器词典的设计与实现. [A].新加坡 1996 年中文电脑国际会议(ICCC"96)论文集[C].

[4] 林杏光等. 现代汉语述语动词机器词典，北京语言学院出版社. 1994.

[5] 刘扬，俞士汶，于江生. CCD 语义知识库的构造研究. 小型微型计算机系统. Aug. 2005. Vol.26. No.8.

[6] 王惠，詹卫东，俞士汶. 现代汉语语义词典的结构及应用. 语言文字应用. Feb. 2006. No.1.

[7] 董振东. 关系：词汇语义的灵魂 .第二届词汇语义学研讨会 北京大学 2001.5 .

[8] 董振东，董强. 知网. http://www.keenage.com.

[9] F. Suchanek, G. Kasneci, and G. Weikum, "YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia," In Proceedings of WWW 07. ACM Press, 2007. Pages: 697–706. WWW 2007/Track: Semantic Web.

[10] Torsten Zesch and Christof Müller and Iryna Gurevych. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. Proceedings of the Conference on Language Resources and Evaluation (LREC), electronic proceedings. Mai (2008).

[11] 张如莹，黄居仁. 中央研究院中英双语知识本体词网 (Sinica BOW)：结合词网，知识本体与领域标记的词汇知识库. in Proceedings of ROCLING 2004(第十六届自然语言与语音处理研讨会,台湾). 2004.

[12] 孙春葵，钟义信. 面向应用领域的语义词典构造研究. 情报学报. Aug.2000. Vol.19 Supplement. Pages:2-3.

[13] 熊忠阳，张鹏招，张玉芳. 基于 χ~2 统计的文本分类特征选择方法的研究. 计算机应用, Feb.2008. Vol.28 No.2.

[14] 程莉，卢正鼎，文坤梅，李娟. 基于语义的模糊匹配探索与应用. 华中科技大学学报（自然科学版）. Feb. 2003. Vol.31 No.2. Pages:1-2.

[15] 夏天. 汉语词语语义相似度计算研究. 计算机工程(Computer Engineering). Mar. 2007. Vol.33 No.6. Pages:1-3.

[16] 黄连恩，王磊，李晓明. 一种基于 LCS 的相似网页检测算法. 北京大学信息科学技术学院 . 网络与信息系统研究所(PKU_CS_NCIS_TR2007012). Dec.2007.