Journal of Computer Applications
计算机应用,2013,33(3):838 -840,844
ISSN 1001-9081
CODEN JYIIDU
2013-03-01
http://www.joca.cn

文章编号:1001 -9081(2013)03 -0838 -03
doi:10.3724/SP.J.1087.2013.00838

协同过滤在中文维基百科类别推荐上的应用

王　静[1,2],何婷婷[1,2*],衣马木艾山·阿布都力克木[3]
(1.华中师范大学 计算机学院,武汉 430079;　2.国家语言资源监测与研究中心 网络媒体语言分中心,武汉 430079;
3.国家数字化学习工程技术研究中心(华中师范大学),武汉 430079)
(*通信作者电子邮箱 tthe@mail.ccnu.edu.cn)

摘　要:针对传统人工编辑导致大量类别信息重复和不规范的问题,提出了应用协同过滤技术为中文维基百科文章自动推荐类别。利用中文维基百科中的四个重要语义特征即链入、链出、链入的类别和链出的类别来表示维基百科文章,得到与目标文章相似的前若干篇文章的所有类别后,通过查询返回的相似度值计算各个类别的权重,选择前面的若干个类别作为推荐结果返回给目标文章。实验结果表明了这四个语义特征能较好地表征一篇维基百科文章,同时也验证了协同过滤方法在中文维基百科自动推荐类别中的有效性。

关键词:协同过滤;中文维基百科;类别推荐;语义特征
中图分类号:TP391.1　**文献标志码**:A

Application of cooperative filtering in categories recommendation of Chinese Wikipedia

WANG Jing[1,2], HE Tingting[1,2*], Yimamu'aishan ABUDOULIKEMU[3]
(1. *School of Computer Science, Central China Normal University, Wuhan Hubei* 430079, *China*;
2. *Network Media Branch, National Language Resources Monitoring and Research Center, Wuhan Hubei* 430079, *China*;
3. *National Engineering Research Center for E-Learning* (*Central China Normal University*), *Wuhan Hubei* 430079, *China*)

Abstract: Collaborative filtering was applied to automatically recommend categories for a Chinese Wikipedia article. Four typical semantic features namely incoming link, outgoing link, incoming link categories and outgoing link categories, were adopted to represent articles. Among all the categories of articles similar to target article, several most similar categories were chosen as the recommendation results to the target article, via calculating the similarity value between them. The experimental results show that the four semantic features have efficient performance in Wikipedia article representation. And the collaborative filtering method is also proved to be effective in recommending proper categories for Chinese Wikipedia articles.

Key words: collaborative filtering; Chinese Wikipedia; category recommendation; semantic feature

0　引言

开放分类是维基百科[1]文章的一个重要信息,每篇维基百科文章可以被志愿者指定一个或多个开放分类用来描述这篇文章的类别信息。传统的人工编辑过程中,由于缺乏对整个维基百科分类系统的大致了解和认识,志愿者在为目标文章编辑类别时常常会遇到许多疑惑和困难,可能会贴上意思相近的不同类别标注,这样最终会造成大量类别信息的重复和不规范。另一方面,当文章被重新编辑修改时其类别信息也需要再次编辑修改。因此,人工标注类别标签是一件耗时而费力的事情。针对此问题,本文提出了应用协同过滤[2]技术为中文维基百科文章自动推荐开放分类。

目前已有许多研究者在维基百科数据集上进行了大量应用研究,例如语义相关度计算、文本分类、信息检索、标注推荐等。文献[3-9]分别利用中英文维基百科的不同特点进行语义相关度计算,验证了维基百科丰富的结构化信息在语义表征上的有效性。文献[10]利用维基百科的类别信息表征文本,引入了类别信息的重要性但还不够全面。文献[11]在中文维基百科中用基于关键词的方法为文本推荐链接标注以达到文本扩充的目的,由于中文维基百科数据库的稀疏性,效果没有英文维基百科好。文献[12]提出用协同标注的方法为英文维基百科推荐类别。他们的方法利用英文维基百科的链入、链出、信息盒和章节名四个语义特征取得了较好的效果,然而中文维基百科与英文维基百科在数量和质量上有较大区别,如何为中文维基百科自动推荐类别仍是一个值得研究的问题。

利用协同过滤为中文维基百科自动推荐类别的主要过程是找到能提供类别信息的与之相似的维基百科文章,目标文章和相似文章都可以通过中文维基百科中广泛存在的四种语义特征来表示,进而进行相似度计算,利用相似度值对从相似文章集中得到的所有类别进行计算打分,选出分值最高的前面若干个类别作为推荐结果返回给目标文章。与其他一些需要获取全文信息的推荐技术相比,这种轻量级推荐方法不需要大量数据和时间进行训练,是即时的。

1　中文维基百科中的语义特征

维基百科的基本单位是文章,或叫条目。如图1所示,是中文维基百科中一个描述实体"华中师范大学"的条目。这

收稿日期:2012-09-18;**修回日期**:2012-11-16。　**基金项目**:国家自然科学基金资助项目(90920005, 61003192);国家语委"十二五"重点项目(ZDI125-1);国家"十二五"科技支撑计划项目(2012BAK24B01);教育部/国家外国专家局高等学校学科创新引智计划项目(B07042);湖北省自然科学基金资助项目(2011CDA034);华中师范大学中央高校基本科研业务费专项资金资助项目(CCNU10A02009, CCNU10C01005)。

作者简介:王静(1989-),女,湖北荆州人,硕士研究生,主要研究方向:自然语言处理;　何婷婷(1964-),女,湖北黄冈人,教授,博士,主要研究方向:自然语言处理、数据库;　衣马木艾山·阿布都力克木(1973-),男(维吾尔族),新疆伊宁人,副教授,博士,主要研究方向:网络信息检索。

《计算机应用》Journal of Computer Applications

篇中文维基百科文章包含了标题、链接、分类以及信息盒等内容。由于维基百科结构的特殊性,可以充分利用其丰富的语义信息而不需要用整个全文信息来表示一篇维基百科文章,对于类别推荐来说,最重要的语义信息应该是包括了链接信息、链接的类别信息以及信息盒等模板信息。然而,由于中文维基百科的发展时间较短,编辑修改条目的志愿者人数也相对较少,所以目前的中文维基百科整体的数量和质量相对于成熟的英文版本而言有较大劣势,并且很多条目缺乏例如信息盒、章节名等模板项,考虑到中文维基百科的特点和不足,一般采用中文维基百科中广泛存在的链接信息以及链接的类别信息作为代表其文章信息的语义特征。

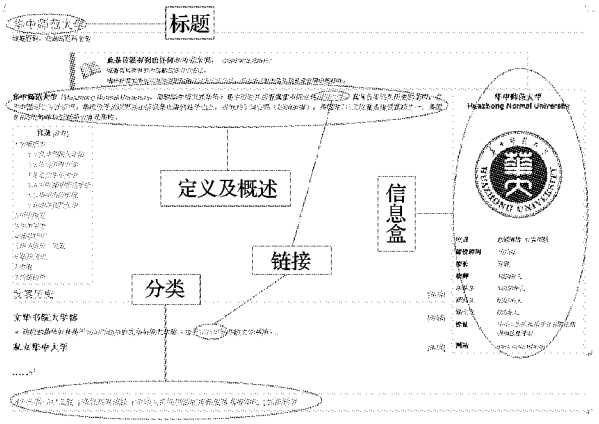

图 1　条目“华中师范大学”

维基百科文章中的链入链出构成了一个隐形的语义网络,用户可以通过点击内部超链接轻松方便地从一篇文章跳转到与它相关的其他文章,因此链接信息能较好地反映两篇维基百科文章之间的语义关系,在相关度计算上是一个重要因素。例如,条目“比尔盖茨”有一个超链接文本“微软公司”可以链接到条目“微软”,同时“微软”也有链接文本可以链接到条目“比尔盖茨”,这两个条目之间有链入链出关系,显然它们之间也是有很大相关性的。

维基百科文章中的内部链接是一个新的条目,同时也有类别信息即链入的类别和链出的类别,由于内部链接通常与目标文章有较高相关度,因此内部链接的类别信息也与目标文章有一定相关性,也可以作为语义特征表示的一个组成部分。

2　类别推荐

协同过滤是一种常见的推荐算法。社会性标注中传统的协同过滤方法利用用户和项目之间的二元关系,首先找到与目标用户兴趣相似的用户,然后根据这些相似用户的兴趣偏好为目标用户推荐项目。

维基百科中的类别不同于传统的分类类别,是一种开放分类,这种由志愿者人为添加的类别标签与社会性标注中用户添加的社会标签相似。因此,受社会性标注中推荐算法的启发,本文考虑到应用协同过滤的方法为维基百科推荐类别。另外,维基百科中已有的类别标注已经足够丰富,可以被重用来标注一篇新的维基百科文章。同时两篇文章如果在上述的语义特征上有很大共性,则它们在类别标注上也会有较大共性。

因此,类别推荐方法描述如下:首先,找到与目标文章在语义信息上相似的文章,得到相似文章集;然后,对相似文章集的所有类别标注通过排序公式计算打分,选择前面若干个分值最大即最合适的类别标注作为推荐结果。

2.1　获取相似文章

采用向量空间模型(Vector Space Model,VSM)中的信息检索方法和文档表示来计算两篇维基百科文章的相似度,选择上述中提到的四种语义特征来表示维基百科文章并进行相似性计算:

1) 链入 I。给定一篇维基百科文章 d,得到所有链入 d 的文章的标题,定义 d 的链入表示 d_I 为 $\{t_{I_1},\cdots,t_{I_{|d_I|}}\}$,两篇文章如果都被同一个条目链入则它们很相似。

2) 链出 O。给定一篇维基百科文章 d,得到所有 d 链出的文章的标题,定义 d 的链出表示 d_O 为 $\{t_{O_1},\cdots,t_{O_{|d_O|}}\}$,两篇文章如果都链出同一个条目则它们很相似。

3) 链入的类别 IC。给定一篇维基百科文章 d,得到 d 所有链入的类别的标题,定义 d 的链入类别表示 d_{IC} 为 $\{t_{IC_1},\cdots,t_{IC_{|d_{IC}|}}\}$,两篇文章如果共享较多的链入类别则它们很相似。

4) 链出的类别 OC。给定一篇维基百科文章 d,得到 d 所有链出的类别的标题,定义 d 的链出类别表示 d_{OC} 为 $\{t_{OC_1},\cdots,t_{OC_{|d_{OC}|}}\}$,两篇文章如果共享较多的链出类别则它们很相似。

构建四个域 d_I、d_O、d_{IC} 和 d_{OC} 形成一个虚拟文档 vd 来表示 d。这四个域的语义特征与相似性计算是很相关的,采用标准的 TF-IDF 模型来计算目标文章 d 和文章 d_i 之间的相似度 $sim(d_i,d)$。对于每一篇用户正在编辑的文章 d,通过其虚拟文档表示以布尔形式作为查询内容检索得到相似结果的排序列表,选择前 n 个相似文章 $d_1,d_2,\cdots,d_n$ 作为相似文章集合 $D(d)$。

另一种最常用的特征是全文信息,然而,由于全文信息的处理需花费较多时间且提高并不明显,而语义特征已经足够丰富用来表示维基百科文章,所以本文放弃使用全文信息而使用语义特征来表示。

2.2　类别排序

第二个步骤是对相似文章集合 $D(d)$ 中的所有候选类别进行排序。定义 f 表示每个类别的分值,对于一个候选类别 c,所有标注了 c 的相似文章形成一个集合 SA_c,$score_{c,d_i}$ 为类别 c 在文章 d_i 中的得分。类别 c 的最终得分 f_c 为相似文章集 $D(d)$ 中的所有文章下的得分之和,如式(1) 所示:

$$f_c = \sum_{d_i \in SA_c} score_{c,d_i} \tag{1}$$

如何计算 $score_{c,d_i}$?直观上越相似的文章提供的类别标注越重要,因此可以把 $sim(d_i,d)$ 作为候选类别的重要性分值,同时对类别重要性加上一个系数 $|D|/rank_{d_i}$,其中 $|D|$ 表示返回的所有相似文章的数量,$rank_{d_i}$ 表示 d_i 在相似文章排序列表集合中的序号。最终 $score_{c,d_i}$ 由式(2) 确定:

$$score_{c,d_i} = sim(d_i,d) \cdot |D|/rank_{d_i} \tag{2}$$

这样做的实际意义是对每篇相似文章赋予了不同的权重,除了考虑它们与目标文章的相似度外,最相似的文章即排在第一位的文章被放大了 $|D|-1$ 倍,系数是 $|D|$,而最不相似的文章系数为 1,即重要性保持不变。

3　实验结果及分析

本文从中文维基百科数据下载页面[13] 下载了原始的数据集,包括条目、类别、链接等数据库表以及包含文章内容的 XML 文件,通过对原始数据进行整理分析,最终将数据存储

Journal of Computer Applications 计算机应用

在 MySQL 数据库系统中,语料中包括了 314 527 个维基百科文章、143 345 个类别和 22 159 921 个内部链接。

整个系统应用 Lucene 作为搜索引擎,每篇维基百科文章由链入、链出、链入的类别和链出的类别四个特征域组成。通过查询索引得到相似文章集。

实验评估采用目标文章在维基百科中原始的类别标注作为基准,将推荐结果与之比较。主要用到的评价指标是信息检索中的三个常用指标,即准确率、召回率和 F 值。

3.1 不同特征的评估

不同的语义特征有不同的文章表征效果。实验中随机抽取了 200 篇文章进行测试,考虑查询结果得到的相似文章的数量,这一过程中将相似文章集的所有类别不经过排序全部返回,不同语义特征表示下的召回率结果如图 2 所示。从图 2可看出链出的分类最能表示维基百科文章。

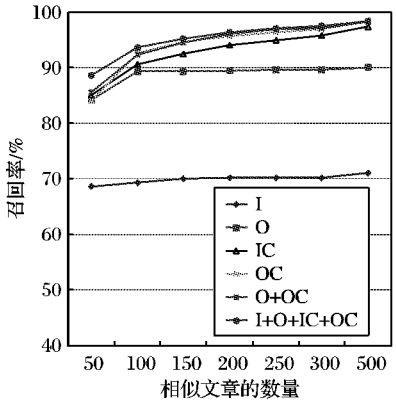

图 2 不同语义特征的比较和组合

由于丰富全面的语义信息,四种语义特征组合的结果是最好的,因此这也被选为下面实验中默认的文章表征方式。对于一篇新的维基百科文章,其链入和链入的类别一般较难获得,幸运的是,链出和链出类别的组合方式也有很好的语义表征效果。同时,相似文章的数量也对召回率有一定影响,综合考虑实验结果和效率,在以下的实验中选择 200 篇相似文章返回。

3.2 最佳推荐个数

不同的类别推荐个数有不同的实验效果,本文在数据集中分别选择人名 100 篇、地名 100 篇、机构名 100 篇和随机 200 篇维基百科文章进行测试。表 1 给出了 1 ~ 10 个类别推荐个数下,不同数据集的准确率和召回率。

表 1 不同推荐个数下的准确率和召回率

推荐个数	人名 100 篇		地名 100 篇		机构名 100 篇		随机 200 篇	
	准确率	召回率	准确率	召回率	准确率	召回率	准确率	召回率
1	0.604	0.144	0.745	0.356	0.578	0.320	0.770	0.443
2	0.542	0.253	0.690	0.638	0.406	0.376	0.589	0.565
3	0.479	0.327	0.624	0.701	0.300	0.398	0.457	0.619
4	0.413	0.364	0.542	0.701	0.249	0.428	0.390	0.635
5	0.333	0.366	0.420	0.743	0.215	0.450	0.338	0.648
6	0.294	0.383	0.358	0.761	0.180	0.460	0.290	0.650
7	0.260	0.390	0.306	0.761	0.160	0.478	0.246	0.665
8	0.240	0.406	0.264	0.775	0.147	0.496	0.216	0.675
9	0.215	0.406	0.231	0.775	0.133	0.500	0.193	0.682
10	0.195	0.410	0.208	0.777	0.119	0.500	0.174	0.686

表 1 显示推荐排序列表的前面几个类别时,准确率很高,这表明排序方法把最合适的类别排在前面,也验证了类别排序的有效可行性。随着推荐类别个数的增加,准确率降低、召回率增大,因此考虑将平衡准确率与召回率的 F 值作为评价指标。图 3 中给出了不同推荐类别个数下的 F 值,由于中文维基百科的特点,人名通常包含出生、去世等年份信息,且类别标注较多较杂,所以推荐效果较差,而相反地名推荐效果较好。综合来看,随机选择 200 篇原始维基百科文章测试,F 值能达到 0.53。实验结果同时表明在中文维基百科中最佳的类别推荐个数是 2,这也与测试文章中平均的类别标注个数为 3 基本相符。

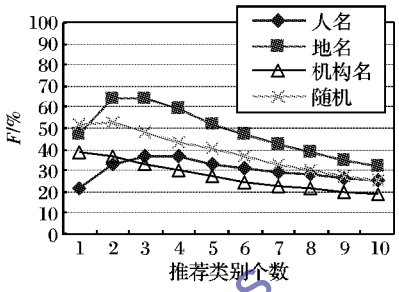

图 3 不同推荐类别个数下的 F 值

4 结语

本文采用了协同过滤的方法通过重用语义特征相似的文章的类别标注自动为中文维基百科文章推荐开放分类,本文方法基于这样一个假设,即与目标文章在语义特征上相似的文章能为其提供相似且合适的协同类别标注。实验结果表明了协同过滤在中文维基百科数据集上的有效性,也可以将此方法应用于其他类似的有相似语义特征的数据集上做测试。

今后的进一步工作包括以下方面:1)通过加入一些其他的语义特征例如信息盒模板、粗体黑字等来精炼维基百科文章的语义表征,提高准确率和实验效果;2)将本文方法应用于其他数据环境并评估效果,可以考虑包含丰富标注信息的社会性标注网站(如豆瓣、知乎等)。

参考文献:

[1] Wkipedia [EB/OL]. [2011-02-06]. http://wikipedia.jaylee.cn/.

[2] VOSS J. Collaborative thesaurus tagging the Wikipedia way[J]. Wikimetrics, 2006, 1(1): 1-7.

[3] STRUBE M, PONZETTO S P. WikiRelate! Computing semantic relatedness using Wikipedia[C]// AAAI '06: Proceedings of the 21st National Conference on Artificial Intelligence. Madison: AAAI Press, 2006: 1419-1424.

[4] MILNE D, WITTEN I H. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links[C]// Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence. Madison: AAAI Press, 2008: 25-30.

[5] TORSTEN Z, CHIRISTOF M, IRYNA G. Using wikitionary for computing semantic relatedness[C]// Proceedings of the 23rd AAAI Conference on Artificial Intelligence. Madison: AAAI Press, 2008: 861-867.

[6] 张红春. 中文维基百科的数据整理和词语间语义相关度计算[D]. 武汉: 华中师范大学, 2011.

[7] LI Y, HUANG K Y, REN F J, *et al*. Wikipedia based semantic related Chinese words exploring and relatedness computing [J]. Journal of Beijing University of Posts and Telecommunications, 2009, 32(3): 109-112.

[8] 盛志超, 陶晓鹏. 基于维基百科的语义相似度计算方法[J]. 计算机工程, 2011, 37(7): 193-195.

(下转第 844 页)

Journal of Computer Applications 计算机应用

	$Item_1$	$Item_2$	$Item_3$	$Item_4$	…	$Item_{30}$
$User_1$	3	3	5	4	…	3
$User_2$	4	2	4	2	…	5
$User_3$	5	1	3	3	…	4
$User_4$	1	4	3	5	…	2
⋮	⋮	⋮	⋮	⋮		⋮
$User_{20}$	5	2	2	3	…	3

图 4 用户-图书评价矩阵

推荐质量的评价标准有多种，本文采用平均绝对偏差（Mean Absolute Error，MAE）对推荐质量进行评价。平均绝对偏差 *MAE* 通过计算用户的预测评分与用户实际评分之间的偏差来度量预测的准确性，*MAE* 越小，推荐质量越高。平均绝对偏差 *MAE* 的计算公式如式（5）所示：

$$MAE = \frac{\sum_{i=1}^{n} |p_i - q_i|}{N} \tag{5}$$

其中，通过预测得到的用户评分集合表示为 $\{p_1, p_2, \cdots, p_n\}$，实际的用户评分集合为 $\{q_1, q_2, \cdots, q_n\}$。

实验中，用户的共同评分项目数 β 由专家给出为 2。θ 为变量，随着 θ 数目的变化能更进一步检验本文方法的准确性。当 θ 值较大时，本文方法能给出更为准确的推荐结果。

3.2 结果分析

将本文改进方法与文献[3－4]方法进行比较。以平均绝对偏差 *MAE* 作为推荐质量的评价标准，比较结果如图 5 所示。

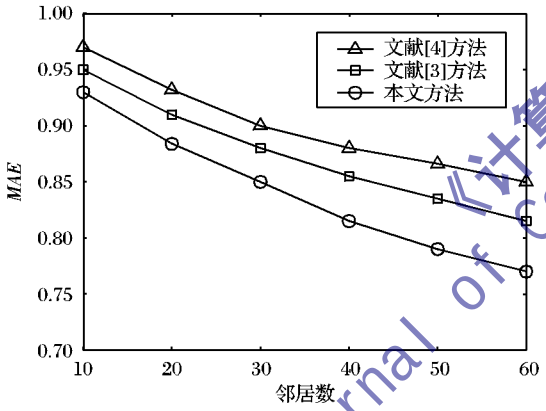

图 5 推荐算法的 *MAE* 比较

由图 5 可以看出，随着用户邻居数目的不断增加，本文方法总是比文献[3－4]方法具有较小的 *MAE* 值，这表明本文给出的方法具有较高的推荐精确度。原因在于加入用户间的信任关系后，目标用户邻居集的构建更为准确，进而使评分预测更加准确。

4 结语

本文利用 SNA 技术分析用户间信任关系对推荐有效性的影响，将用户间的信任关系融入到协同推荐研究中，对协同推荐方法进行了改进。实验结果表明，本文方法能够有效地解决协同推荐中的数据稀疏和冷启动问题，较大程度地提升了推荐的质量，具有良好的应用前景。

参考文献：

[1] 马宏伟，张光卫，李鹏. 协同过滤推荐算法综述[J]. 小型微型计算机系统，2009，30(7)：1282－1288.

[2] 李春，朱珍敏，高晓芳. 基于邻居决策的协同过滤推荐算法[J]. 计算机工程，2010，36(13)：34－36.

[3] 罗辛，欧阳元新，熊璋，等. 通过相似度支持度优化基于 *K* 近邻的协同过滤算法[J]. 计算机学报，2010，33(8)：1437－1445.

[4] 邓爱林，朱扬勇，施伯乐. 基于项目评分预测的协同过滤算法推荐算法[J]. 软件学报，2003，14(9)：1621－1628.

[5] DESHPANDE M，KARPIS G. Item-based top-*n* recommendation algorithms[J]. ACM Transactions on Information Systems，2004，22(1)：143－177.

[6] 王岚，翟正军. 基于时间加权的协同过滤算法[J]. 计算机应用，2007，27(9)：2302－2326.

[7] LIU J，WANG Q P，FANG K. An optimized collaborative filtering approach combining with item based prediction[C]// Proceedings of the 11th International Conference on Computer Supported Cooperative Work in Design. Piscataway，NJ：IEEE Press，2007：158－160.

[8] 刘军. 社会网络分析导论[M]. 北京：社会科学文献出版社，2004.

[9] FREEMAN L C . The development of social network analysis：a study in sociology of science [M]. Vancouver：Empirical Press，2004.

[10] 罗家德. 社会网络分析讲义[M]. 北京：社会科学文献出版社，2010.

[11] AHN H J. A new similarity measure for collaborative filtering to alleviate the new user cold starting problem[J]. Information Science，2008，178(1)：37－51.

[12] GRANDISON T，SLOMAN M. A survey of trust in Internet applications[J]. IEEE Communications Survey and Tutorials，2000，4(3)：2－16.

（上接第 840 页）

[9] 刘军，姚天昉. 基于 Wikipedia 的语义相关度计算[J]. 计算机工程，2010，36(19)：42－46.

[10] 王锦，王会珍，张俐. 基于维基百科类别的文本特征表示[J]. 中文信息学报，2011，25(2)：27－31.

[11] 杨柳. 基于中文维基百科的文本扩充[D]. 武汉：华中师范大学，2011.

[12] WANG Y，WANG H F，ZHU H P，*et al*. Exploit semantic information for category annotation recommendation in Wikipedia[C]// Natural Language Processing and Information Systems，LNCS 4592. Berlin：Springer，2007：48－60.

[13] 中文维基百科资源[EB/OL]. [2011-02-06]. http://dumps.wikimedia.org/zhwiki/.

[14] 王刚. 自动抽取维基百科文本中的语义关系[D]. 上海：上海交通大学，2011.

[15] 李赟. 基于中文维基百科的语义知识挖掘相关研究[D]. 北京：北京邮电大学，2009.

[16] 张海粟，马大明，邓智龙. 基于维基百科的语义知识库及其构建方法研究[J]. 计算机应用研究，2011，28(8)：2807－2811.

[17] 熊忠阳，史艳，张玉芳. 基于维基百科和网页分块的主题爬行策略[J]. 计算机应用，2011，31(12)：3264－3267.

[18] CARMEL D，ROITMAN H，ZWERDLING N. Enhancing cluster labeling using wikipedia[C]// SIGIR '09：Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York：ACM Press，2009：139－146.

[19] XU Y，JONES G J，WANG B. Query dependent pseudo-relevance feedback based on Wikipedia[C]// SIGIR '09：Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York：ACM Press，2009：59－66.

《计算机应用》Journal of Computer Applications