# 中文维基百科的结构化信息抽取及词语相关度计算*

张红春[1,2]，何婷婷[1,2]，涂新辉[1,2]，周琨峰[1,2]

[1]华中师范大学 计算机科学系，湖北 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，湖北 武汉 430079

E-mail: zhclk@yahoo.com.cn

**摘 要**：本文首先从中文维基百科官方所提供的基本数据中抽取整理出多种结构化信息；接着，对维基百科的知识组织形式进行了抽取架构，实现了一套开放的框架接口，方便了用户对这些信息的获取和使用；在此基础上，进行了词语间语义相关度计算的实验，并把实验的结果与传统的经典方法进行了对比，证明了利用中文维基百科进行语义研究的可行性。

**关键词**：语义相关度；中文维基百科；结构化信息

# Extracting Structured Information from the Chinese Wikipedia and Measuring Relatedness Between Words

Zhang Hong-chun[1,2], He Ting-ting[1,2], Tu Xin-hui[1,2], Zhou Kun-feng[1,2]

[1]Department of Computer Science, HuaZhong Normal University, Wuhan 430079

[2]Network Media Branch, National Language Resources Monitoring and Research Center, Wuhan 430079

E-mail: zhclk@yahoo.com.cn

**Abstract:** In this paper, after extracting structured information from the Chinese Wikipedia, we abstract the objects which the Wikipedia uses to organize the knowledge, and implement an open source framework. Finally, we compute semantic relatedness between words using the structured information, and compare its performance with the traditional classical method. The experimental results prove that it is feasible to use Wikipedia for semantic research.

**Keywords:** semantic relatedness; Chinese Wikipedia; structured information

## 1 引言

为了提高计算机的智能化程度，在自然语言处理的过程中，加入语义知识的理解是非常必要的。随着日益增长的信息处理需求，如何从海量的语料资源中自动地获取丰富的语义知识，以及如何有效地利用这些语义知识来实现对文本语义的理解，已成为一个重要的研究课题。

维基百科作为一个以开放和用户协作编辑为特点的 Web 2.0 知识系统，具有知识面覆盖度广，结构化程度高，信息更新速度快等优点，其中蕴涵有丰富的语义知识，是目前众多学者进行语义知识抽取所青睐的语料数据资源。近几年来，国外的许多学者专家以英文维基百科作为语料库、语义知识库，从不同的角度抽取语义知识进行研究，取得了很多突破性的成果。Michael Struble 和 Simone Paolo Ponzetto 最先利用维基百科进行了语义相关度的研究[1]。他们把在 WordNet 知识库上效果比较好的一些经典算法移植到维基百科的分类图中，并使用三个测试数据集 M&C、R&G 和 WS-353 分别做了实验。结果表明，在大数据集上，维基百科的计算结果要远好于 WordNet。Zesch et al.利用德语版的维基百科和 GermaNet 进行了类似的实验研究，得出了一致的结果[3]。Gabrilovich 和 Markovitch 提出了显示语义分析算法(Explicit Semantic Analysis)，简称 ESA[4]。他们的主要思

* 项目资助：国家自然科学基金重大研究计划课题(90920005)；国家自然科学基金(61003192)；973 国家重点基础研究发展计划课题(2007CB310804)；教育部哲学社会科学研究重大课题攻关项目(08JZD0032)；教育部/国家外国专家局高等学校学科创新引智计划课题(B07042)；湖北省自然科学基金计划项目(2009CDB145)；武汉市晨光计划项目(201050231067)；华中师范大学中央高校基本科研业务费项目(CCNU10A02009, CCNU10C01005)

想是采取一种基于质心的策略，把文本的语义隐射到一个由维基百科概念所形成的带有权重的高维空间向量中，再利用简单的向量乘法计算两者之间的语义相关度。Milne 则只分析使用了文档中出现的内链接，他认为链接应该代表了更强烈的语义信息[5]。

目前，在国内，基于中文维基百科的研究还比较少，也没有一些开源的工具和数据可以使用。在词语间语义相关度计算方面，李赟博士提出的综合多条关联路径的算法[6]，综合考虑了分类图和文档图中两节点间路径的条数、每条路径的长度、图中节点间的不同关联程度等多个特征。

本文首先从中文维基百科官方所提供的一些半结构化的基本数据中抽取整理出内链接信息，分类图和锚文本等多种类型的结构化信息；接着，对维基百科的知识组织形式进行了抽象架构，实现了一套开放的框架接口，方便了用户对这些信息的获取和使用；在此基础上，本文综合利用内链接和锚文本信息做了词语间语义相关度的计算实验，实验结果表明了使用中文维基百科进行语义研究的可行性。

## 2 维基百科的结构化信息抽取

维基百科的内容是以网页的形式提供给用户使用，每个网页代表其描述的一个实体或概念，简称条目，具有唯一的整数标号 page_id。每个条目包括标题和正文两个部分，在正文部分常含有大量指向其他条目的内链接。并且，每个条目都归属于一个或多个分类，而每个分类又可以拥有一个或多个子分类，这样所有的分类就组成了一个有向无环的层次结构。维基百科的这种独特的知识组织形式使得其中含有丰富的结构化语义信息，但是，维基百科的官方仅提供一些半结构化的基本数据文件的备份，很多有用的结构化语义信息和数据并不能直接地获取和使用。

为此，本文首先从中文维基百科官方网站下载了 2010 年 08 月 29 日这个日期版本的备份数据文件，接着，在综合阅读了国外大量的基于维基百科的研究工作之后，主要集中对三个方面的数据进行了抽取整理：条目的正文内容、链接和链接的锚文本。其中，对条目正文内容的处理主要包括中文繁体转简体、文本中噪音的过滤（如：模板、表格、外连接等）、文本的索引等几个方面。对链接的处理又包括对分类链接的处理和对内链接的处理两个方面，前者主要是进行分类图的建立、分类深度的计算以及条目与分类的从属关系提取；后者主要是统计条目间的内链接、重定向链接、消歧义链接。最后，结合文本内容和内链接信息，对内链接的锚文本的使用情况进行了统计。在实现的过程中，本文把抽取的结构化数据都按表存储在数据库中，并对各个重要字段建立了索引。经过一系列地处理之后，本文得到了七张数据表，包含了对中文维基百科中锚文本、内链接和分类的统计情况，数据表的结构和规模如表 1 所示。

为了让用户在后续的工作中能够更直观地把握，更方便地获取和使用这些结构化的信息和数据。本文首先分析了维基百科中条目的不同作用，进而把所有的条目为了六种类型：普通的解释性条目，分类条目，消歧义条目，重定向条目，消歧义义项条目和锚文本义项条目；接着，从整体上对维基百科的知识组织形式进行了抽象架构，针对每一类抽象实体，总结并实现了获取其相应的结构化信息的方法，最终实现了一套开发的框架接口。例如：对于普通的解释性条目，本文实现了获取其链入链接、链出链接、所属分类、正文的锚文本数据、正文内容等信息的接口；对于分类条目，实现了获取其子分类、父分类、所有属于该分类的普通条目、该分类在分类图中的深度等信息的接口；对于重定向条目，实现了获取其重定向到的条目等信息的接口；对于消歧义条目，实现了获取其正文中所列举出的所有义项等信息的接口；对于锚文本，实现了获取该锚文本所链接到的所有的不同条目等信息的接口。利用这套接口，用户可以通过简单的对象初始化和方法调用就可以获取和使用这些结构化的信息。

表1 中文维基百科整理后数据统计

| 数据表名 | 规模 | 数据表意义 | 字段名 | 字段的意义 |
|---|---|---|---|---|
| Categorycategorylinks_id | 140397 | 分类图 | cl_parent | 父类别的 page_id |
| | | | cl_child | 子类别的 page_id |
| Anchor | 1059798 | 锚文本 | anchor_text | 链接的锚文本 |
| | | | anchor_to | 链接指向条目的 page_id |
| | | | anchor_count | 链接出现的总次数 |
| Categorydepth_id | 77064 | 类别在分类图中的深度 | category_id | 类别的 page_id |
| | | | category_depth | 类别的深度 |
| Categorylinks_id | 1552739 | 条目到所属类别的链接 | cl_from | 条目的 page_id |
| | | | cl_to | 条目所属类别的 page_id |
| Disambiguation | 7641 | 消歧义条目与所包含的义项 | dg_from | 消歧义页的 page_id |
| | | | dg_to | 义项的 page_id |
| | | | index | 是第几个义项 |
| | | | scope | 义项的简短介绍 |
| Pagelinks_id | 22159921 | 普通条目之间的内链接 | pl_from | 链接的起始条目 page_id |
| | | | pl_to | 链接的目标条目 page_id |
| Redirect_id | 317913 | 重定向 | rd_from | 重定向的起始条目 page_id |
| | | | rd_to | 重定向的目标条目 page_id |

# 3 基于中文维基百科的词语间语义相关度计算

本文把计算词语间语义相关度的方法分为两个大的步骤：①把两个词语 A 和 B 的当前含义分别映射到维基百科条目 C 和 D 上；②计算条目 C 和 D 的语义相关度值 sim(C,D)作为 A 和 B 的语义相关度。

## 3.1 词语到维基百科条目的映射

在自然语言中，一词多义的现象极为普遍，然而，当需要计算两个词语的相关度时，它们的当前含义是唯一确定的。因此，词语的映射也包含两个步骤：①找出词语所有的含义；②从所有的含义中确定词语的当前意义。

在维基百科中，最简单的查找一个词语的所有义项条目的方法就是在不考虑标题补充说明的前提下，判断该词语与哪些条目标题相等，把所有符合条件的条目作为该词语的义项集合，但是，这种方法在很多时候并不能找出所有的义项。而中文消歧义条目的数目又非常少，许多多义词语都没有与之对应的消歧义条目，利用消歧义条目也只能完成少部分的工作。分析发现，维基百科条目之间含有大量的内链接，而这些链接的锚文本要么是所指向条目的标题，要么是所指向条目的标题的别名。因此，采用查询词语与锚文本相等的方法，会获得更好的结果。

确定词语当前含义的过程，本文考虑了两个方面的因素：一个是普遍性，一个是相关性。普遍性是指一个词语的某个义项被人们所熟知的程度，这个值可以通过锚文本的使用情况来近似获得，比如：词语 A 作为锚文本，指向的条目有 $< c_1, c_2, \ldots, c_n >$，每种链接在整个维基百科中出现的次数分别为 $< k_1, k_2, \ldots, k_n >$，则词语 A 常被用作义项 $c_i$ 的比率为：

$$r_i = \frac{k_i}{\sum_{j=1}^{n} k_j} \tag{1}$$

相关性是指词语 A 的所有含义与词语 B 的所有含义，两两配对计算相关度，哪一对义项的得分越高，则认为这对义项更可能是词语 A 和词语 B 的当前含义。最后，通过普遍性和相关性值的线性加权，选取总得分最高的那个义项对的相关性值作为词语 A 和词语 B 的相当度。

### 3.2 条目间的语义相关度计算

条目间的链接包含着丰富的语义信息，可以理解为阅读该条目所需要的背景知识。分类也非常重要，不同的作者在表达同一个意思时，所使用的链接条目可能差异很大，但是，这些不同的链接条目所属于的分类差异却相对小了很多。并且，一个条目可以属于多个类别，加入分类信息，可以减少仅考虑链接时可能遇到的数据稀疏性问题。因此，本文综合考虑了条目的链入链接、链出链接、链入链接所属类别和链出链接所属类别四个方面的特征。在实现过程中，先分别计算两个条目在这四种特征空间上的相关度后，再通过线性加权的方式，得到最终的相关度值。

设有两个条目 $w_1$ 和 $w_2$，链接到 $w_1$ 和 $w_2$ 的不同条目的 page_id 所组成的集合分别为 $S_{w1\_linkin}$ 和 $S_{w2\_linkin}$。合并这两个集合就可以形成链入链接的特征空间 $FS_{w1\_w2\_linkin}$，可以形式化的表示为：

$$FS_{w1\_w2\_linkin} =< w\_in_1, w\_in_2, \ldots, w\_in_n > \tag{2}$$

其中，$w\_in_i$ 表示一个不同的条目的 page_id，$n$ 表示合并之后不同条目的个数。此时，条目 $w_i$ 在特征空间 $FS_{w1\_w2\_linkin}$ 中可以表示为：

$$V_{wi} =< tf\_wi(w\_in_1), tf\_wi(w\_in_2), \ldots, tf\_wi(w\_in_n) > \tag{3}$$

其中，$tf\_wi(w\_in_i)$ 表示条目 $w\_in_i$ 链接到条目 $w_i$ 的次数。有了条目 $w_1$ 和 $w_2$ 的向量表示，就可以计算余弦值得到两个条目在链入链接的特征上的相关度 $R_{linkin}$。

进一步，设 $FS_{w1\_w2\_linkin}$ 中每个条目 $w\_in_j (1 \leqslant j \leqslant n)$ 所属类别分别为集合 $C_{wj\_in\_cat}$，则由 $FS_{w1\_w2\_linkin}$ 中所有条目的类别集合 $C_{wj\_in\_cat} (1 \leqslant j \leqslant n)$ 合并就形成了链入链接的类别特征空间 $FC_{w1\_w2\_linkin}$。此时，条目 $w_i$ 在特征空间 $FC_{w1\_w2\_linkin}$ 上可以表示为：

$$VC_{wi} =< tf\_wi\_c(w\_in\_cat_1), tf\_wi\_c(w\_in\_cat_2), \ldots, tf\_wi\_c(w\_in\_cat_m) > \tag{4}$$

其中，$w\_in\_cat_i$ 表示每个不同的类别 page_id，

$$tf\_wi\_c(w\_in\_cat_j) = \sum_{k=1}^{n} tf\_wi\_tmp(w\_in_k) \tag{5}$$

$$tf\_wi\_tmp(w\_in_k) = \begin{cases} tf\_wi(w\_in_k) & w\_in\_cat_j \in C_{wk\_in\_cat} \\ 0 & other \end{cases} \tag{6}$$

通过计算余弦值，又可以得到两个条目在链入链接所属类别特征上的相关度 $R_{linkin\_cat}$。

按照类似的方法，还可以得到链出链接和链出链接所属类别特征上的相关度 $R_{linkout}$ 和 $R_{linkout\_cat}$。最后，条目 $w_1$ 和 $w_2$ 的语义相关度计算公式为：

$$R_{w1\_w2} = \alpha * \frac{R_{linkout} + R_{linkin}}{2} + (1-\alpha) * \frac{R_{linkout\_cat} + R_{linkin\_cat}}{2} \tag{7}$$

## 4 实验结果和分析

本文随机选取了多个领域的 30 对词语组成了测试数据集，让 10 位不同专业背景的研究生各种独立对每对词语进行相关度打分，分值在 0~10 直接，0 表示完全不相关，10 表示等价。然后，对每对词语的 10 个打分计算平均值，来得到数据集中每对词语的人工相关度判断结果。

在具体的评测过程中，从多个方面做了对比实验：1）单独考虑四个特征的相关度值；2）分别只考虑链接和类别的相关度值；3）综合考虑四个特征的相关度值 $R_{w1\_w2}$；4）与文献[8]方法的对比；5）调整 $\alpha$ 的值，来对比链接和分类的重要性。

附表 1 中显示了 $\alpha = 0.5$ 时的计算结果，所有的计算结果都是四舍五入后，保留了 5 位小数。可能由于少数词对人工打分的原因，使得基于 HowNet 的方法的相关系数显得比较低，但是从整体的分数分布上，可以得出以下结论：1）HowNet 的计算结果有很多分数是相同的，且不同的词对间分数的分布波动性不大，而维基百科的分数分布显得更合理，这从某种程度上说明了维基百科克服了传统知识库数据稀疏的问题；2）对比只考虑链接的计算和只考虑类别的计算，可以看出对于质量较好的条目，只考虑链接的计算准确性已经很高，对于一般的条目，加入类别信息，可以较小数据的稀疏性问题；3）综合考虑链接和类别信息，可以适当的减少因引入目前还不完善的分类图所可能带来的语义偏移。当 $\alpha$ 从 0.1~0.9 变化时，$R_{w1\_w2}$ 与人工打分的相关系数逐渐递减，这也从某种程度上反映了维基百科分类系统的不完善性，与统计的结果一致。

## 5 结论和展望

本文首先抽取整理了中文维基百科结构化信息；接着，对维基百科的知识结构进行了抽象处理，实现了一套开放的框架接口；在此基础上，进行了词语语义相关度计算的实验，进而证明了利用中文维基百科进行语义研究的可行性。在后续的工作中，可以更加深入的对维基百科中其他的语义知识进行研究运用。

## 参 考 文 献


[1] Michael Strube and Simon Paolo Ponzetto. WikiRelate! Computing semantic relatedness using Wikipedia. In Proceedings of AAAI, pages 1419-1424, 2006.

[2] Simone Paolo Ponzetto, Michael Strube.: Knowledge Derived From Wikipedia For Computing Semantic Relatedness. Journal of Artificial Intelligence Research 30, 181-212 (2007).

[3] Torsten Zesch and Christof Müller and Iryna Gurevych. UsingWiktionary for Computing Semantic Relatedness. In Proceedings of AAAI, pages (861-867), 2008.

[4] Evgeniy Gabrilovich, Shaul Markovitch.: Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. Proceedings of IJCAI, 1606-1611.

[5] David Milne.: Computing Semantic Relatedness using Wikipedia Link Structure. Proceedings of the New Zealand Computer Science Research Student conference (NZCSRSC'07), Hamilton, New Zealand.

[6] 李赟. 基于中文维基百科的语义知识挖掘相关研究: [博士学位论文]. 北京邮电大学, 2009.

[7] http://zh.wikipedia.org/

[8] 刘群，李素建. 基于《知网》的词汇语义相似度计算，第三届汉语词汇语义学研讨会，2002. 5.

[9] Pu Wang, Carlotta Domeniconi.: Building Semantic Kernels for Text Classification using Wikipedia. In KDD '08, New York, NY, USA. ACM.

附表 1　语义相关度计算结果

| 词语 1 | 词语 2 | 人工打分 | HowNet | 链入链出 | 链入链出类别 | $R_{w1\_w2}$ |
|---|---|---|---|---|---|---|
| DNA | 遗传 | 7.6 | 0.11815 | 0.02234 | 0.28787 | 0.15510 |
| DNA | 魔术师 | 0.5 | 0.19302 | 0.00357 | 0.05283 | 0.02820 |
| DNA | 基因 | 9.73 | 0.86 | 0.05549 | 0.25242 | 0.15395 |
| 不明飞行物 | 行星 | 2.7 | 0.21053 | 0.01454 | 0.25447 | 0.13450 |
| 不明飞行物 | 飞碟 | 8.9 | 1 | 1 | 1 | 1 |
| 不明飞行物 | 外星人 | 6.2 | 0.15094 | 0.01697 | 0.15235 | 0.08466 |
| 操作系统 | 微软 | 4.8 | 0.21747 | 0.08326 | 0.40508 | 0.24417 |
| 操作系统 | 内存 | 2 | 0.17833 | 0.03724 | 0.24783 | 0.14253 |
| 操作系统 | 电子邮件 | 1.3 | 0.29619 | 0.02029 | 0.14251 | 0.08140 |
| 股票 | 股市 | 5.8 | 0.20932 | 0.10874 | 0.49823 | 0.30348 |
| 股票 | 公司 | 3.3 | 0.19302 | 0.01827 | 0.21225 | 0.11526 |
| 股票 | 风险 | 6.95 | 0.19302 | 0.01201 | 0.19084 | 0.10142 |
| 单亲家庭 | 心理 | 3.8 | 0.27391 | 0 | 0.08526 | 0.04263 |
| 单亲家庭 | 压力 | 4 | 0.09926 | 0 | 0.06520 | 0.03260 |
| 单亲家庭 | 家庭 | 4.7 | 1 | 0.02478 | 0.21163 | 0.11820 |
| 人工智能 | 哲学 | 0.6 | 1 | 0.03206 | 0.30875 | 0.17040 |
| 人工智能 | 自动化 | 4.3 | 0.13704 | 0.02012 | 0.29379 | 0.15695 |
| 人工智能 | 机器人 | 5.2 | 0.16667 | 0.03763 | 0.34102 | 0.18932 |
| 宇宙飞船 | 太空 | 4.6 | 0.18605 | 0.06813 | 0.30567 | 0.18690 |
| 宇宙飞船 | 宇航员 | 6.6 | 0.17833 | 0.04565 | 0.32068 | 0.18316 |
| 宇宙飞船 | 火箭 | 6.4 | 0.29609 | 0.03064 | 0.28285 | 0.15674 |
| 广义相对论 | 黑洞 | 6.9 | 0.18605 | 0.17324 | 0.62431 | 0.39877 |
| 广义相对论 | 相对论 | 6.1 | 1 | 0.13245 | 0.54902 | 0.34074 |
| 广义相对论 | 宇宙 | 5.8 | 0.18605 | 0.07749 | 0.3788 | 0.22814 |
| 华尔街 | 金融 | 5.3 | 0.30333 | 0.00446 | 0.19022 | 0.09733 |
| 华尔街 | 投资者 | 7.4 | 0.22418 | 0 | 0.12703 | 0.06351 |
| 华尔街 | 中国 | 3.6 | 0.58 | 0.00889 | 0.2745 | 0.14169 |
| 测谎器 | 瞳孔 | 2.5 | 0.16364 | 0 | 0.03853 | 0.01926 |
| 测谎器 | 心理学 | 2.5 | 0.15630 | 0 | 0.07439 | 0.03720 |
| 测谎器 | 引渡 | 0.3 | 0.13704 | 0 | 0.02691 | 0.01346 |
| 相关度 | | | 0.20028 | 0.404012 | 0.533424 | 0.49503 |