基于网络百科全书的中文关联语义知识获取*

杨柳[1,2]，何婷婷[1,2]，涂新辉[3]

1 华中师范大学计算机科学系 武汉 430049

2 国家语言资源监测与研究中心网络媒体分中心 武汉 430049

3 国家数字化学习工程技术研究中心 武汉 430079

yangliu721@yahoo.com.cn tthe@mail.ccnu.edu.cn tuxinhui@163.com

摘要：本文提出了一种用语义标签、语义指纹来表示关联语义知识的形式化方法，其中语义标签指代语义中的一个知识单元（也即概念），语义指纹是对语义标签所指代概念的描述，由语义标签的相关词语及其关联度共同组成。本文提出了一种从网络百科全书获取中文关联语义知识的方法，通过该方法获得语义标签的相关词群，利用网络百科全书中的内部链接和开放分类信息计算每个相关词语与语义标签的关联度。通过与人的判断进行比较，说明了本文提出的计算语义标签与相关词语关联度的方法的有效性。

关键字：百科全书，语义知识，语义关联度，语义标签，语义指纹

Obtaining Related Semantic Knowledge from Online Encyclopedia

Liu Yang[1,2] Tingting He[1,2] Xinhui Tu[3]

1 Department of Computer Science and Technology, Huazhong Normal University, Wuhan, China

2 Network Media Branch of National Language Resources Monitoring and Research Center, Wuhan, China

3 Engineering & Research Center for Information Technology on Education, Huazhong Normal University, Wuhan, China

yangliu721@yahoo.com.cn tthe@mail.ccnu.edu.cn tuxinhui @163.com

Abstract: This paper uses semantic label and semantic fingerprint to represent the related semantic knowledge and proposes a method to obtain related semantic knowledge from an online encyclopedia. Semantic label represents a concept, which can be a word or phrase, and is a basic semantic unit in natural language. Semantic fingerprint is consisted of the following pairs: semantic related term, its semantic relatedness. We obtain the semantic related terms of each semantic label and compute the semantic relatedness by analyzing the inner hyperlinks and open category information in the encyclopedia. By comparing our results with human judgments, we prove that our relatedness computing method is effective.

Keywords: encyclopedia; semantic knowledge; semantic relatedness; semantic label; semantic fingerprint

1 引言

万维网的发展使得人类拥有的文本信息资源越来越多，人们迫切需要机器能自动地从海量文本中获取目标信息。自然语言处理领域的一些子任务，诸如文本分类、信息检索、自动文摘和自动问答等都是通过计算机来解决人们对特定文本信息的需求。由于互联网上文本信息大都是以人类的自然语言出现，而人类的语言中蕴含着丰富的语义知识，因此，为了增加机器理解自然语言

* 基金项目：国家自然科学基金重大研究计划(No. 90920005)，国家自然科学基金(No. 60773167)，国家十一五科技支撑计划课题(No. 2006BAK11B03)，973 国家重点基础研究发展计划(No. 2007CB310804)，教育部/国家外国专家局高等学校学科创新引智计划(No. B07042)，湖北省自然科学基金计划项目资(No. 2009CDB145)，武汉市晨光计划项目资助(No. 201050231067)

的能力，在自然语言处理过程中适度地引入语义知识便十分必要。

人类语言的复杂性，使得语义知识的内涵十分丰富，词法、句法、概念的分类结构、词语之间的相似度、词语之间的相关度等都可以属于语义知识的范畴。中文语义知识的来源有两种，一种是人工构建的知识库，如Hownet[1]，另一种则是大规模的真实文本，包括互联网上的海量文本、各种离线文本集合（如各种规模的语料库），各种网络百科全书（如中文维基百科[2]、百度百科[3]、互动百科[4]等）。

国内著名的知识库 Hownet 是通过义原、概念以及义原和概念之间的关系来反映中文语义知识。这种人工构建的知识库一般是由语言学家或领域专家标定而成，知识库的质量很高；但是人工标注耗时耗力，且语言的发展过程中新词新义不断出现，人工构建的知识库很难及时收录各种新词新义。大规模的文本集合是人类自然语言的真实写照，能反映语言中丰富的语义知识，尤其是词语之间的共现规律能在一定程度上反映词语之间的语义关联度。但是互联网上的文本大都是以无结构的形式出现，这给机器自动获取语义知识带来了一定的困难。

网络百科全书是目前存在于互联网上的一种人人可编辑的在线百科全书，它通过分类体系和各种超级链接将大量的概念组织起来，使得概念与概念之间具有一定的关联性，这种关联是语义知识中的一个重要范畴，也是本文研究的内容。网络百科全书有以下显著特点：任何一个可以使用互联网的人都可以在遵循一定规范的情况下在网络百科全书中创建和编辑词条，因此网络百科全书词条的数目远远超过传统依靠少数人力或部分专家构建的知识库中的词条数。另外，网络媒体和各种网络通讯工具的发展，使得新事物和新词的传播速度以及其为人们所接纳的速度也随之加快，网络百科全书的人人可编辑性保证了其会及时收录社会发展过程中不断出现的新事物和新词。进一步地，网络百科全书中的一个词条往往是由多个用户共同编辑、修改而成，因此该词条的综合信息从一定程度上反映的是多个用户的对词条理解，也即群体智能。基于以上发现，本文旨在研究从网络百科全书中获取语义知识。

互联网上中文网络百科全书主要有以下三个：中文维基百科、百度百科和互动百科。与传统百科全书一样，在网络百科全书中，每一个词条都对应有词条解释页面。但是，许多词语在不同的语境中会有不同的含义，比如苹果，可能表示一种水果或者一个叫苹果的公司。中文维基百科和互动百科为这些多义词建立了相应的消歧义页面，即在这个页面中列出该多义词所有可能的含义，用户根据自己的需要点击相应含义的超链接即可进入特定含义下的词条解释页面。也就是说，中文维基百科和互动百科为多义词的不同含义建立了不同的页面。但是百度百科中却没有类似的处理，一个多义词的解释页面同时包含它的多个含义，这给机器获得多义词特定语境下的语义知识带来了一定困难。

由于各种原因，目前中文维基百科的许多词条的原始版本都是以繁体编写，虽然维基百科页面上可以实现繁简转换，但是许多领域词语和国外人名仍然不能得到很好的处理。

表 1 给出了三个网络百科全书的主要不同之处（其中词条数是于 2010 年 1 月 13 日获取得到）。基于以上发现以及各百科全书的特点，本文拟采用互动百科作为中文关联语义知识来源。

---

[1]董振东, 董强（1999），“知网”, http://www.keenage.com

[2]中文维基百科.http://zh.wikipedia.org/zh-cn/Wikipedia:%E9%A6%96%E9%A1%B5

[3]百度百科.http://baike.baidu.com

[4]互动百科.http://www.hudong.com.

表 1. 中文维基百科、百度百科和互动百科的主要差别

	词条数	是否有消歧义页	页面以繁体/简体编写
中文维基百科	290,379	有	许多文章的原始版本是繁体编写
百度百科	1,959,284	无	简体
互动百科	4,357,226	有	简体

互动百科的每个具有特定含义的词条都对应一个解释页面。解释页面中包含有词条名、摘要、目录、正文、开放分类、用户添加的相关词条和其他信息。正文是对该词条的解释，其中的内容与该词条紧密相关。一个词条可以属于多个开放分类，用户可以为一个词条添加多个相关词条。通过该词条解释正文里提到的词语、用户添加的该词条相关词条的超链接，可以链接到互动百科中其他词条的解释页面。

2 关联语义知识的形式化表示

美国心理学家哈里•洛拉尼在他的书中提到[1]：“如果要记住任何新信息，就必须使这条信息与你已经知道或记住的信息联系起来。联系，与记忆紧密相关，指的就是将两种或两种以上的事物捆绑在一起，或者在它们之间建立起关系。”同理，人们遇到一个新概念（知识）时，往往是将存储在人脑中已有的概念与新碰到的概念建立起一定的联系，从而达到理解和记忆新概念的目的。这种概念与概念之间的关系，是语义知识中很重要的一部分。例如，当我们遇到一个概念“基因”（假设我们事先对这个概念的内涵并不了解），同时看到对“基因”的解释语句“基因是指携带有遗传信息的DNA序列，是控制性状的基本遗传单位”时，我们必定在脑海中建立起了“基因-遗传信息”，“基因-DNA”，“基因-遗传单位”等等这样的关系对。人类习得新的知识必定是建立在以往知识经验的基础上，并将新知识与过往经验建立起一定的关联。

因此，本文旨在研究通过建立一种形式化的表示关联语义知识的方法来模拟人类习得知识的过程。本文通过词汇之间的关联关系来形式化地描述这种关联语义知识，将概念作为语义知识的一个基本单元，用语义标签来指代，用语义指纹来刻画其语义。通过语义标签和语义指纹来表示自然语言中的关联语义知识，其具体形式为：

语义标签：语义指纹；

其中，语义标签代表一个概念，对应于语言中的词或短语，它是自然语言中表达某个语义知识的基本单元。但是语义标签并不完全等同于语义知识单元或概念，有时一个词在不同的语境中有完全不同的语义，对应于两个不同的知识单元，这样两个知识单元可能共用一个标签，也即一个语义标签可能指代多个语义知识单元，例如词语“苹果”在不同的语境中可能对应一种水果或一个公司，也即语义标签“苹果”指代了多个语义知识单元，在这种情况下，语义标签所指代的语义知识则由该语义标签和其语义指纹共同确定。

语义指纹是对语义标签所指代的概念的语义描述，用概念的相关词群及每个词对语义指纹的贡献度来刻画。贡献度也表达了语义指纹中的一个词语与语义标签（也是一个词语）的关联度。

表 2 中给出了两个语义标签以及其语义指纹，表中的数字（0 到 1 之间）表示了词语对语义指纹的贡献度，数值越大表明越相关。由于苹果这个语义标签所指代的概念有多个，因此我们根据它所指代的不同概念分成了两行展示。

表 2. 语义标签及其语义指纹示例

语义标签	语义指纹（相关词语及其贡献度）
DNA	(脱氧核糖核酸 1.0),(基因 0.927), (生命科学 0.838),(蛋白质 0.812), (细胞质 0.761) ……
苹果[水果]	(维生素 C 0.752),(营养 0.706),(梨 0.661),(植物 0.646),(苹果果醋 0.639).....
苹果[公司]	(个人电脑 0.798),(电脑 0.791),(手提电脑 0.748),(iPhone 0.676),(乔布斯 0.502).....

3 基于网络百科全书的中文关联语义知识获取

语义知识的获取有多种途径，互联网上海量文本以及各种离线文本（如各种规模的语料库）中的词语共现规律都能在一定程度上反映词语之间的语义关联，另外，网络百科全书中词条与词条之间的链接关系等信息也能反映出词条之间的语义关联，本文提出了一种从网络百科全书中获取关联语义知识的方法。

3.1 语义标签的获取

一般来说，百科全书中的一个词条即代表了一个语义知识单元，但互动百科中，并非所有的词条都能作为语义标签，比如有些词条的解释页面过于短小，那么这个页面所能提供的语义信息就十分有限。另外，如果一个词条的解释页面中提到的并以内部链接形式出现的其他词条也很少时，则表明该词条与其他词条的语义关联很弱。这些词语通常是一些生僻且不常用的词语。本文的第四部分给出对这些词条进行筛选的一些方法。

3.2 语义指纹的获取

获取语义标签的语义指纹也即要先获取该语义标签的相关词群，然后计算每个相关词对语义指纹的贡献度，也就是每个相关词与语义标签的关联程度。

1）相关词群的获取：在互动百科中，一个词条的解释正文中提到的其他词条有一些会以内部链接的形式出现，用户在浏览词条解释正文时点击正文当中的内部链接即可以链接到其他词条的解释页面，例如，互动百科中词条 *DNA* 的解释正文中有这样的语句“DNA 又称脱氧核糖核酸，是染色体的主要化学成分，同时也是基因组成的，有时被称为'遗传微粒'”。这段话中提到的“*脱氧核糖核酸*”、“*染色体*”、“*基因*”等词都以内部链接的形式出现。我们认为这些词条是与被解释的词条在语义上是相关的，并将一个词条（语义标签）的内部链接词条记为 *Inner*。

另外，在互动百科中，用户都会对每个词条添加其相关词条。显然，这些词条也是与被解释的词条在语义上有关联的，我们将其记为 *UserRelate*。

2）相关词语贡献度的计算：在互动百科中，一个词条可以属于多个开放分类，如词条“DNA”，它所属于的开放分类为：分子生物学、分子生物物理学、基因、生物化学、生物医学工程、生物学、生物物理学、遗传学、遗传学术语。这些类别信息共同反映了这个词条的语义知识，进一步地，由于该词条解释正文中的内部链接词条以及用户添加的该词条的相关词条是与该词条在语义上有关联的，所以，该词条解释正文中内部链接词条所属于的开放分类以及用户添加的该词条的相关词条所属于的开放分类也在一定程度上贡献了该词条的语义知识。

我们将互动百科中一个词条（也即语义标签）l自身属于的开放分类集合C_{self}、该词条解释正

文中所有内部链接词条所属于的开放分类的集合C_{Inner}与用户添加的该词条的相关词条所属于的开放分类的集合$C_{UserRelate}$的并集定义为词条l的开放分类语义知识集合SC_l。当两个词语的开放分类语义知识集合中大部分开放分类都是一样的时候，那么这两个词语必定在语义上有很强的相关性。因此，相关词语贡献度的计算可以通过比较语义标签l的开放分类语义知识集合SC_l与其相关词语w_i的开放分类语义知识集合SC_{wi}而得到。

对于上文提到的语义标签“*DNA*”，它的开放分类语义知识集合SC_{DNA}={(生物化学, 49), (分子生物学, 47), (遗传学术语, 43), (医学名词, 34), (生物医学工程, 23), (生物学, 22), (医学术语，21), (基本物理概念, 20), (物理学, 18), (物理化学, 16), ... }，其中圆括号内的数字代表了该开放分类出现的频次。

接下来，在计算语义标签l与相关词语w_i之间的语义关联度时，首先定义一个开放分类语义知识向量v，开放分类语义知识向量的维数n=$|SC_l \cup SC_{wi}|$，每一维代表一个开放分类，记为c_p。语义标签l的开放分类语义知识向量v_l在p维上的值即为l的开放分类语义知识集合SC_l中对应开放分类c_p出现的频次。对相关词w_i的开放分类语义知识向量v_{wi}也有同样的定义。

因此相关词语w_i与语义标签l的语义关联度$r_i = (v_l \cdot v_{wi}) / (|v_l| \times v_{wi}|)$。

对于上述提到的语义标签“*DNA*”，用该计算方法得到其语义指纹结果见表 3。

表 3 语义标签“*DNA*”的语义指纹

语义标签	语义指纹
DNA	(脱氧核糖核酸 1.000),(基因 0.927),(线粒体,0.903),(染色体 0.893),(基因工程 0.883),(核苷酸 0.864),(RNA 0.851),(核糖核酸 0.851),(原核细胞 0.849),(遗传 0.845),(生命科学 0.838),(细菌 0.837),(细胞核 0.836),(分子生物学 0.835),(蛋白质 0.812),(病毒 0.793).....

4 实验

作为实验，通过直接访问互动百科的页面，我们选取了互动百科的部分词条，另外，为了筛选掉其中一部分非常生僻，使用频率非常低的词条，我们又根据搜狗互联网词库[5]筛选出了 8333 个词条作为语义知识库中的语义标签。

4.1 实验过程

为了保证计算语义标签与相关词语语义相关度的准确性，我们筛选掉了那些内部链接和用户添加的相关词条链接总数小于 10 的不完善词条，通过这样的筛选，我们得到的有效语义标签个数为 7964 个。同样的，在通过互动百科得到一个语义标签的相关词群后，我们也筛选掉了那些在互动百科中的解释页面内容过短，而且链接到其他词条的链接也很少的相关词。

筛选后，我们得到了这 7964 个语义标签的相关词群以及其相关词语语义标签的语义关联度。经过统计，我们为每个语义标签获得相关词语的个数平均为 27. 4 个。

4.2 结果分析

[5]搜狗实验室资料下载-互联网词库. http://www.sogou.com/labs/dl/w.html.

为了评价机器计算词语之间相关度结果的好坏，国外学者普遍采用的是将机器得到的结果与人的判断进行比较[2][3][4]，这其中“人的判断”普遍用的Finkelstein等人的词语集合（The WordSimilarity-353 Test Collection）[5]，该词语集合包含了353对词语，每对词语之间的关联度是10个不同的人对该词语对的关联度判断的平均值。目前，国内没有类似的中文词语集合，为了验证采用本文中的方法计算语义标签与相关词语的关联度的有效性，我们首先随机选取了10个语义标签，然后随机选出了每个语义标签中的三个相关词语，让10个不同专业背景的研究生各自独立地对这10*3对词语的语义关联度进行判断，判断的值在10-0之间，10表示完全相关，0表示完全不相关。

表4为这10*3对词语的语义关联度的人工判断结果的平均值，以及采用本文提出的方法计算得到的语义关联度。为了将本方法计算的结果与人工判断的结果进行比较，采用了Pearson相关系数来计算机器结果与人工结果的相关性。表中最后一行为采用Pearson相关系数得到的本文计算出的词语语义关联度与人工判断的相关性，可以看到其值为0.67，说明本文提出的计算语义标签与相关词语的语义关联度的有效性。

5 结论与展望

本文提出了一种用语义标签、语义指纹来显式表示中文关联语义知识的形式化方法，并研究了一种利用网络百科全书来获取语义知识的方法，作为实验，我们得到了7964个语义标签的语义指纹，平均每个语义标签的相关词语个数为27.4个。利用Pearson相关系数判断相关度计算方法得到的结果与人工判断的相关值为0.67，说明了本方法计算语义标签与其相关词语的语义关联度的有效性。

在接下来的工作中，我们将进一步研究如何从大规模的文本集合中感知概念间的语义关联，进而获取语义指纹。同时研究如何将所获取的语义知识应用到自然语言处理的子任务中以提高机器理解自然语言的能力。

参 考 文 献

[1]哈里·洛拉尼（著）；徐建萍（译）. 超级记忆力训练Ⅰ+Ⅱ+Ⅲ+Ⅳ. 西安：陕西师范大学出版社，p19, 2009.8.

[2]Michael Strube and Simon Paolo Ponzetto. WikiRelate! Computing semantic relatedness using Wikipedia. In *AAAI'06*, Boston, MA, 2006.

[3] Gabrilovich, E. and S. Markovitch. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. *Proceedings of IJCAI*, 1606-1611.

[4] S. P. Ponzetto and M. Strube. Knowledge derived from wikipedia for computing semantic relatedness. *Journal of AI Research (JAIR)*, 30:181–212, 2007.

[5] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin, "Placing Search in Context: The Concept Revisited", *ACM Transactions on Information Systems*, 20(1):116-131, January 2002

表 4. 语义相关度计算结果

词语 1	词语 2	人工评价	互动百科计算结果*10
DNA	遗传	8.3	8.46
DNA	魔术师	0.5	2.79
DNA	基因	9.73	9.27
UFO	行星	3.7	3.69
UFO	飞碟	8.9	10
UFO	外星人	7.5	6.99
操作系统	微软	6.7	7.63
操作系统	内存	3.5	5.71
操作系统	电子邮件	1.9	5.75
测谎器	瞳孔	6.6	6.63
测谎器	生理心理学	8.4	7.08
测谎器	引渡	3	3.71
单亲家庭	心理	6	5.59
单亲家庭	压力	5.2	6.1
单亲家庭	家庭	6.25	3.94
股票	股市	8.6	9.81
股票	公司	6.2	2.51
股票	风险	6.95	6.1
广义相对论	黑洞	6.9	9.22
广义相对论	相对论	8.2	8.67
广义相对论	宇宙	5.8	8.17
华尔街	金融	8.4	7.31
华尔街	投资	7.85	7.05
华尔街	中国	3.6	1.58
人工智能	哲学	1.5	5.66
人工智能	自动化	7.95	5.9
人工智能	机器人	7.8	6.11
宇宙飞船	太空	8.35	6.39
宇宙飞船	宇航员	6.9	6.86
宇宙飞船	火箭	6.9	5.52
相关度			0. 67