基于本体语义树的主题空间向量模型①

卢承山

(武汉理工大学 计算机科学与技术学院，武汉 430063)

摘 要：在传统检索模型的基础上，结合本体的概念，提出一种基于本体语义树的主题空间向量模型。该模型能够用语义概念树描述一个主题，与传统基于关键词描述主题的方法不同，它能够描述概念之间的简单语义关系。在此基础上，给出 HTML 页面内容与主题相关度的计算方法。在分析 URL 的相关度时，不仅分析链接锚文本与主题相关度，还结合了改进的 PageRank 算法来分析链接的相关度。只有当链接相关度达不到给定的阀值时才会去下载链接对应的页面。这样的 URL 相关度计算方法可以大大减少不必要的计算开销, 又可以充分地利用锚文本和链接重要度信息。最后还对那些不确定是否与主题相关的网页进行内容相关度计算，进而最终确定是否应该采集此网页。

关键词：本体；概念树；主题网络；锚文本；主题相关度

Thematic VSM Based on Ontology Semantic Tree

LU Cheng-Shan

(School of Computer Science and Technology, Wuhan University of Technology, Wuhan 430063, China)

Abstract: Based on the traditional search model, combining the concept of ontology, this paper proposes a thematic network crawling model based on ontology semantic tree. Unlike the traditional keyword-based subject description methods, the model can describe a subject with semantic concept tree with which it is simple to describe the semantic relationships between concepts. On this basis, the paper presents a method to calculate the relevance of HTML pages and the topic. When analyzing the relevance of URL, it does not only analyze the relevance of link anchor text and the topic, but also analyzes the relevance of the link with an improved PageRank algorithm. Only when the relevance does not reach a given threshold will it download the page corresponding to the URL. This calculation method can greatly reduce unnecessary computational overhead, and make fully use of anchor text and link importance of information. Finally, it calculates the relevance of a web page which is not sure whether it is related to the topic, and ultimately determines whether this page should be collected or not.

Key words: ontology; semantic tree; thematic network; anchor text; degree subject

1 引言

随着网络技术的飞速发展 ，互联网已经成为全球最大的信息载体。人们对网络信息的需求越来越个性化，如何从海量的 WEB 数据源中寻找用户需要的数据，是目前 WEB 信息检索领域的研究热点。主题爬虫[1]是限定爬虫在一定主题领域范围内下载网页的爬虫。因此，以何种策略来选择爬虫的爬行路径,是决定能否采集到所需要信息的关键问题。目前,主题爬虫的爬行策略主要有两种：基于网页链接结构的搜索策略和基于内容评价的搜索策略。前者通过分析网页之间的相互链接关系来确定网页的重要性，进而决定链接访问顺序。该方法虽然考虑了链接结构和网页之间的链接关系但忽略了页面与主题的相关性, 会出现搜索的主题漂移问题。后者起源于文本检索中对文本相似度的评价, 能够准确的评价网页内容与主题的相关性，然而忽略了链接之间存在的结构信息，因而在预

① 收稿时间:2011-02-01;收到修改稿时间:2011-03-14

测链接网页价值方面存在一些不足。

综合考虑以上两种情况，本文结合以上两方面的优点，多粒度来评价网页与主题的相关度。一方面分析链接的相关度，另一方面在链接无法确定情况下，对网页的内容进行分析。主要思想是通过分析网页链接和链接文件与主题相关度来排除与主题完全不相关的链接，选取部分与主题相关的链接，如果通过分析链接和锚文件无法判断与主题相关性的，最后下载此网页，分析网页内容与主题相关性。

2 基于本体语义树的信息采集模型

2.1 主题爬虫的整体结构

主题爬虫是建立在普通爬虫基础之上的，通过在网页抓取的整个过程中增加新模块来实现领域化的提取。这些模块主要包括主题确立模块、链接预测模块，主题相关性分析模块等。那么要实现主题爬虫，主要是在通用的搜索引擎的基础上实现主题领域集的定义和获取，链接的预测和判断、对下载下来的内容进行主题相关性的计算。图 1 给出了主题爬虫整体结构图。

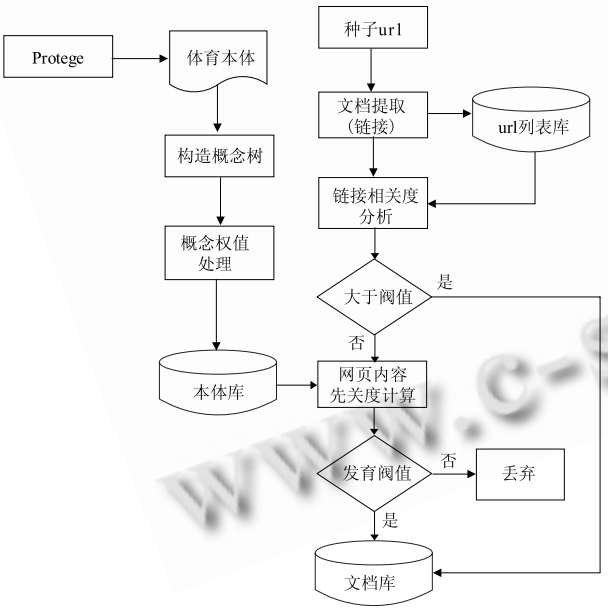

图 1 主体爬虫结构图

2.2 本体学习和构建

主题爬行采用关键词集合描述爬行主题，然后根据网页内容和主题词集来计算相关度，最后决定抓取相应的网页。这种实现方式虽然简单可行，但是存在一词多义或者一义多词现象，造成遗漏少数相关网页和搜索精度不高等问题。显然采用关键词集合来判断相关性的主题爬行不够准确，所以本文提出了一种基于语义判断相关的主题爬行。在语义描述方面，利用本体[2,3]充当主题爬行的领域知识库角色，利用它对客观事物的概念和关系进行详细的描述，从而使采集的信息不偏离主题。本体作为知识库的特点为实现信息检索，智能搜索等领域提供了一种完善的解决方案。因此，为了提高查全率和查准率，本文将本体作为主题爬行的知识库指导爬虫对 web 网页进行抓取。

本体的构建过程是一项艰巨而复杂的工作，每个领域都存在各自特点的概念、关系和约束，并且数量一般都很巨大。近几年来，本体构建工具出现了很多，并且也日趋成熟，如 Ontolingua，OntoSaurus、WebOnto、Protégé等[4,5]。这些工具都提供了良好的图形界面和一定的检查机制。用户不需要关注描述本体语言的细节，只需要集中到组织本体内容上。但是，目前本体构建大多数是基于人工的方式构建，在领域专家和技术人员的参与下构建领域本体。目前本体构建方法主要有：IDEF-5 方法、骨架发、七步法、TOVE 法[6]。本文主要是参考现有的领域本体。参考已经存在的领域本体目的是为了让本体重用，能够减轻构建的工作量，吸取以前构建较为成功的本体内容，应用到本文中。

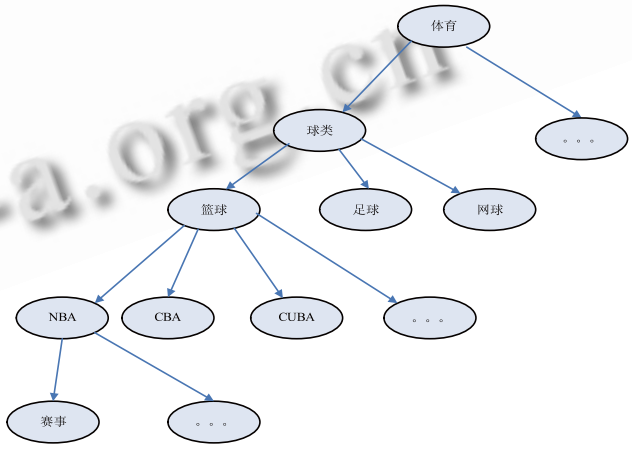

图 2 本体结构图

本文本体构建采取的是 Protégé[7,8]，Protege 是生成和编辑本体与知识库的可扩展、跨平台且开放源码的开发环境，目前已在 30 多个国家得到广泛的使用和推广。Protege 从本体中产生特定领域的知识获取工具和应用，允许领域专家通过生成或改进可复用的本体和解决问题的方法来建立基于知识的系统。它是目前

最容易使用的本体编辑工具，能够定义类和类层次、属性关系和属性-值约束，以及类和属性之间的关系。该工具框架通过配置可扩展 Plug-in 体系方便地集成第三方开发的其他插件。

以上是利用 Protégé构建的简单的本体结构图，目前在很多领域都有成熟的本体，我们可以在目前成熟的本体上加以改进即可得到较为完善的本体。

2.3 构造概念树及生成相应权值

本文使用本体来确定领域主题，为了计算网页与主题相关度，需要对本体中所有的概念[9]和属性进行量化，形成｛概念，权值｝键值对。人为的为每个概念赋值存在很多缺点，如赋值过程中，概念越多工作量就越大；每个领域的本体各有各的特点，人为决定概念的权值的大小也缺少一定的客观依据等。所以本文提出了一种自动根据本体之间的关系，将本体解析为语义树，然后根据语义树自动为每个概念赋值。该方法需要遵循以下规则：

规则 1：在本体的概念之间有继承关系，同义关系和属性关系三种，继承关系表示概念是父类和子类的关系；同义关系表示概念间具有同等地位，可以彼此代替；属性关系表示一个概念是为了描述另一个概念的关系。

规则 2：继承关系和属性关系中，他们是子类和父类，具体和抽象的关系，所以子类比父类在主题领域中更加具体，权值更加大，属性描述比它所描述的概念的权值也更大，属性和概念间的关系如公式（1）；同义关系中，由于概念之间彼地位同等重要，他们的权值应该是相同的。

$$p=\begin{cases}1 & \text{概念}\\ 0.8 & \text{描述概念的属性}\end{cases} \tag{1}$$

基于以上两个原则，使用 Jena 工具解析领域本体，将其构建成一棵概念树，根据概念的关系确定树的层次，并按照定义规则计算概念的权值。通过领域本体的概念树计算得出的权值，组成所有概念和相应的权值的键值对列表｛概念，权值｝，为计算网页内容相关度做准备。

$$W_{(node)}=init+\frac{L}{2H}*P \tag{2}$$

其中，$W_{(node)}$ 表示本体概念的权重，init 表示概念树的根节点的初始值，H 表示概念树的高度，L 表示概念在树中所处的层。P 用来区分概念和属性。

2.4 链接相关度计算

PageRank 算法：假设一个 Web 漫游者每次按统一概率 d 在当前网页中挑选下一步要访问的链接，而当遇到一个没有链出的网页时，漫游者以一个小概率[1 d]跳到任意的一个页面，则在任意时间点，漫游者位于页面 u 的概率为 PR(u)

$$PR(url)=(1-d)+d\sum_{i=1}^{n}\frac{PR(Ti)}{C(Ti)} \tag{3}$$

式中：PR(url)——网页的 PageRank 值，即网页的重要性；PR(Ti)——链接到的网页的 PageRank 值；C(Ti)——Ti 的链出数量；d——阻尼系数，0<d<1，通常取值为 0.85。本文也使用了 PageRank 算法。

当然完全根据链接来计算也是不够准确的，因为链接本身也可能含有与主题相关的丰富的信息，如链接的 title，锚文件，链接本身。

例如：http://www.whut.org/Sports/Basketball 这个 URL 包含的内容就很可能是关于 Basketball 的。其计算公式如(3):

$$\Phi_{uh}(url)=\begin{cases}1 & \text{如果 url 中包含主题词}\\ 0 & \text{否则}\end{cases} \tag{4}$$

如果 URL 的 Text 中包含某个主题词，则这个 URL 所指向的页面也很可能是跟这个主题词密切相关的，例如：

体育

其包含的内容就很可能是关于“体育”，其计算公式如(4):

$$\Phi_{teh}(url)=\begin{cases}1 & \text{如果url的Text包含主题词}\\ 0 & \text{否则}\end{cases} \tag{5}$$

同 URL 启发式算法类似，根据这个公式计算的值 Φ_{teh}，如果为 1，则这个链接所指向的页面与主题相关的准确性很高，但计算的值Φteh 如果为 0，这个链接所指向的页面与主题相关的准确性并不高。

如果一个链接中的 Title 包含某个主题词，则这个 URL 所指向的页面同样也很可能是跟这个主题词密切相关的，例如：

<A href=”http://sports.sina.com.cn/k/2011-01-04/12585393652.shtml “ title=” YaoMing”>

YaoMing is very good!

</A>

在这个 URL 中，title 包含的内容 YaoMing 就很可能是关于这个 URL 所指向的页面的内容，其计算公式

如公式(5):

$$\Phi_{tih}(url)=\begin{cases}1 & \text{如果url的Title包含主题词}\\ 0 & \text{否则}\end{cases} \quad (6)$$

最后，将所有的扩展元数据平均加权综合在一起，就得到扩展元数据启发式算法公式如公式

$$\Phi_{amh}(url)=\frac{A*\Phi_{uh}+B*\Phi_{teh}+C*\Phi_{tih}}{3} \quad (7)$$

最后得到的总的链接的相关度：

$$\Phi_{w}(url)=a\Phi_{amh}(url)+bPR(url) \quad (8)$$

其中 a,b 为调节链接相互关系和链接内容重要性的系数，其中 a+b=1,一般认为如果链接内容中包含了与主题相关，则此 URL 很大程度上是与主题相关的，所以这里取 a=0.6，b=0.4。

2.5 网页内容相关度计算

相关度[10]计算模块是主题爬虫中最重要的功能模块之一，是判断主题相关度的核心模块。网页内容相关度的计算主要是通过流行的向量空间模型来实现，通过主题概念的向量空间和网页中提取的向量空间进行计算而得到相似度。主题概念的向量主要通过以上本体构建，以及通过本体构建概念树，生成相应的权值｛概念，权值｝。而网页中提取的向量空间，需要通过本体中的概念去与网页中切分后的提取词进行匹配得到。并统计出现的次数和出现在网页中的位置，通过计算得到相应网页特征向量的权值。

主题相关度的计算是采用向量空间模型[11]算法。把关键词的个数 n 作为向量空间的维数，每个关键词的权值 W_i 作为每一维分量的大小，则主题用向量表示为：$a=(w_1,w_2,\cdots,w_n)$，其中 i=l,2,…,n。对前边从页面中挖掘纯文本信息进行分词，描述一片文档的内容，一般以词或者词组的形式表示，即文档是特征的集合，可以形式化的表示为 d=（t1，t2，…,tn）。然而单纯地将一篇文档直接分词得到特征词集合，然后统计文档集在主题集合中出现的主题词和次数很难精确的描述一篇文档与主题的相关性，因为它缺乏了语义信息。据观察大部分文档在描述某个主题的时候，是以句子为单位来描述的，如果一个句子中多词出现主题集中的概念，证明此文档越可能是与主题相关的文档。基于以上的发现，我们分词的时候以句号为单位，将分词的结果进行分组，如果在一句话中多次出现主题集合中的概念，我们需要对结果进行一个加权。这里需要对以前的空间向量模型进行一个改进。

如果一句话中多次出现了主题集中的概念，说明在这句话中出现的几个主题词对文档很重要，那么我们需要对出现的主题词的概率做一个加权。如果一个句子中出现与主题相关的主题词的次数为 n（n>0），那么将这次词出现的个数乘以 2n-1

$$f=\begin{cases}\mathrm{d} & \text{否则}\\ \mathrm{d}+2^{n-1} & \text{当前词的句子中出现多个主题集中的词}\end{cases} \quad (9)$$

其中，d 为当前词出现的个数，n 为当前词所在的句子中在主题集中出现的词的个数。

$$W_d=\frac{f}{m} \quad (10)$$

其中，W_d 表示当前词在文档中的权值，而 m 表示当前文档中出现次数最多的词的次数。

根据以上两个公式，我们就可以计算所有主题集概念在文档中的权值，那么用空间向量表示为 b=(wd1,wd2,…,wdn)，i=1，2，3，...，n，用两个向量的夹角的余弦表示页面的主题相关度，公式 11 如下：

$$\cos\langle a,b\rangle=\frac{(a,b)}{|a||b|}=\frac{|x_1w_1^2+x_2w_2^2+\cdots+x_nw_n^2|}{\sqrt{w_1^2+w_2^2+\cdots w_n^2}\sqrt{x_1^2w_1^2+x_2^2w_2^2+\cdots+x_n^2w_n^2}} \quad (11)$$

3 实验步骤以及结果分析

3.1 实验步骤

本实验的具体步骤按照图 1 所示，主要概括如下：

（1）构建本体，到权威本体网站获取某一领域的本体，我们这里下载的是体育方面的本体，下载得到的本体为 owl 描述的本体。

（2）对本体进行更新，下载得到的权威本体可能存在不足，还要针对领域对它进行完善，阅读相关文献丰富本体含义，这是个不断更新的过程。然后利用 protégé重新得到最新的领域本体。

（3）根据本体来构建本体语义树，按照以上方法为本体各个概念和描述赋权值，将会得到｛概念，权值｝组。

（4）然后给定初始 URL 种子，爬取链接，锚文件等相关信息。

（5）计算链接相关度，如果达到一定阈值，爬取网页存储，如果没有达到，下载网页内容，计算网页内容与主题的相关性，最后决定是丢弃还是存储。

3.2 实验结果分析

本文主要是检验采集到的信息是否是与主题相关的，检验基于本体语义树的信息采集系统的查全率和

查准率是否比传统的基于关键词的要准确。本实验分别开启了 10，30，50，70，90，110 个线程在初始种子为“ http://sports.sina.com.cn/”进行爬行。所用网络为教育网，考虑到网络不稳定，各线程的数据是通过 10 次求平均值所得。

表 1 爬行数据表

线程数	时间	爬行过的页面	采集到的页面	随机抽取 100 页面, 相关的网页数目
10	60	4867	2423	97
30	60	6245	2976	95
50	60	7216	3408	94
70	60	9438	4212	94
90	60	8613	3924	95
110	60	7189	3371	93

通过表 1 可以看出，在 10-70 的各个跨度间，爬行速度与采集的速度逐步加快，当线程到达 90 以后爬行速度减慢。主要是由于本系统采取的是多线程，有涉及到内存利用率、同步问题，线程越多会出现争用同一资源，相互等待的状态。我们同时发现我们随机抽取的 100 网页中平均大约有 95 个网页是与主题相关的，基本上保证了采集网页的准确率。

下面对基于关键词（KW）和基于本体语义树（OC_T）两种方法进行对比：

表 2 关键词和本体语义的比较

描述方法	采集时间(分钟)	主体相关页面	查准率	平均下载时间
KW	48	3261	49.34	0.8832
OC_T	63	4134	84.73	0.914

通过以上两种方法的比较，基于关键词的查准率比较低，因为它只是简单的列出了相关主体的概念，然后用概念与页面的相关词进行关键词匹配，来进行相关度的计算，这样忽略了与概念相关的上下文概念，所以它的下载范围比较窄，查准率相对比较低。基于本体语义树利用了关键词及上下文关键词稽核，扩充了主体范围，所以查准率比基于关键词的提高了较多。

对已采集到的少数不准确的网页经过分析，基本上是由于在链接分析的过程中误认为有些链接是与主题相关的，产生这种错误的原因主要是因为有些链接的 title 或者锚文件是与主题相关，但是网页的内容却不是与主题相关的。

4 总结与展望

本文提出了一种基于本体语义树的主题集的确立和主题相关度的算法。该主题集的确立算法能通过本体转化为语义树，比较全面的收集领域的信息，能够通过该主题集准确的判断抓取的信息是否偏离主题，是否与主题相关，能够解决一词多义问题。但是实验中还存在一些问题，本体的构建可能还不够完善，需要不断更新本体才能保持采集的信息的准确度。

参考文献

1 周立柱,林玲.聚焦爬虫技术研究综述.计算机应用,2005,25(9):1965-1969.

2 李文杰,赵岩.基于本体结构的概念间寓意相似度算法.计算机工程,2010,23.

3 张帆,钟金宏,黄玲.改进的领域本体概念相似度计算方法.计算机应用,2010,23.

4 万捷,腾至阳.本体论在基于内容信息检索中的应用.计算机工程,2003,(4):67-71.

5 Ehring M, Maedche A. Ontology–focused crawling of web documents. Proceedings of the 2003 ACM symposium on Applied computing, 2003,1(3):624-626.

6 刘维群,李元臣.Web 信息的语义概念检索.现代情报,2005,7:74-76.

7 Protégé. Protégé HomePage. Http://protege.stanford.edu/, 2005.

8 FIPA.FIPA 98 Specification Part 12: Ontology Service. Http://www.fipa.org/spees/fipa00006/OC00006A.html,1988.

9 Lin D, Pantel P. Concept discovery from text. Proc. of Conference on Computational Linguistics 2002. Taipei, 2002. 577-583.

10 贾雪峰,王建新,齐建东.基于领域本体的智能检索系统.计算机工程,2010,23.

11 雷景生,林冬雪,符浅浅.基于改进向量空间模型的 Web 信息检索技术研究.计算机工程,2005,31(1):14-16.