一种形式概念分析的单页面内容概念树构建方法*

何伟[1,2]，李霜[2]
（1. 中国科学技术信息研究所，北京 100038；2. 怀化学院，怀化 418008）

摘要：在海量信息环境下，网络信息的杂乱，严重影响用户的有效利用。基于页面内容的可用性评估成为研究人员关注的热点。提出了一种形式概念分析的单页面内容概念树构建方法，该方法以单个页面作为切入点，以页面段落为对象，段落中的动名词为属性，构造概念格，利用剪枝命名转换规则，构建概念树。实验结果表明，该方法构建的概念树对页面内容重构以及页面分类合理性判断有较好的效果，可作为Web内容可用性评估的一种手段。

关键词：形式概念分析；概念格；概念树构建
中图分类号：G250.73
DOI：10.3772/j.issn.1673-2286.2015.07.005

1 引言

随着网络信息技术的快速发展，网络信息量成指数级增长，面对海量的页面信息，用户面临信息“富余”的窘迫。网络信息多、杂、乱，严重影响用户对网络信息的有效利用，又回到了信息贫乏时代。为了改变这一现状，使网络信息更好的为用户服务，Web可用性评估逐渐成为国内外研究人员关注的热点。Web可用性评估可从物理可用性和Web内容可用性两方面进行。Web内容体现了Web设计者构建Web的目的，更具主观性，对Web内容的评估更难以进行。然而随着语义技术的发展，使得通过挖掘页面语义信息对Web内容的可用性评估成为可能。通过构建概念树来反映Web内容间的语义信息，能较好的描述页面内容间的联系，可在一定程度上解决Web内容可用性评估的问题。

2 相关研究

2.1 研究现状

页面内容概念树是完成网站内容可用性评估的一个重要手段和前提条件。近年来，许多研究人员提出了各种概念树构建方法。王战军等人提出了一种评估指标的概念树构建方法，他使用层次分析法将评估指标按归类方式进行抽象，建立树状的层次结构，并用概念树对评估指标的相关性进行分析[1]。宣士斌利用概念相容性实现了一种概念树的生成算法，可动态的根据增加属性值调整概念的层次结构[2]。毛宇梅提出了基于链接结构的Web概念树构建方法，该方法通过网页间的超链接并获取链接锚点来完成Web概念树的构建[3]。孙亚琳等人则提出了一种基于主题词表和FCA的网页语

* 本研究得到湖南省教育厅一般项目“基于语义分析的Web可用性评估研究”（编号：13C716）和中国博士后科学基金项目“基于叙词表语义关系的智能检索模型研究”（编号：2014M550791）资助。

义概念树构建方法，他们首先将整个Web站点的网页作为对象，网页特征项为属性，构建形式背景；其次使用主题词表计算词语间相似性，对形式背景进行规范化和简化处理；最后采用FCA构建Web概念树[4]。而杨小平等人则利用概念树对Web语义结构进行评估，结果表明使用概念树对Web内容可用性评估是一种非常有效的方法[5]。

上述的概念树构建方法基本上是以整个Web站点作为研究对象，鲜少有以单个页面作为分析对象的报道。基于此，本文以单个页面作为切入点，以页面内容的段落作为研究对象，提出了一种形式概念分析的单页面内容概念树构建方法。本方法以页面段落为对象，以段落中的动、名词为特征属性，通过形式概念分析来构建概念树，并将其应用到页面分类的正确性判断以及页面内容是否与标题相符。实验证明本文方法所构造的页面内容概念树对页面分类以及内容重构合理性是比较有效的，可进一步改善页面内容的可用性设计。

2.2 相关知识概述

（1）概念树

概念树起源于数据库，是数据库中各属性值和概念依据抽象程度形成的层次结构，是实现和表示概念的一种概括语义描述[6]。随着语义技术的发展，概念树逐渐成为语义网络的核心，定义了特定领域的概念及概念间关系。概念树的分层树形结构特征表述了根节点（顶层）概念是最通用概念，概念所在层次越低，概念越具体。上层概念包含下层概念，若将其反映在网页分类中，则可描述成属于下层概念分类的网页，一定也归类与其上层概念。因此，对单个页面构建内容概念树，可直观的描述页面所传达的内容以及所属的分类。

（2）形式概念分析

形式概念分析（Formal Concept Analysis FCA）于1982年由Willer首次提出[7]，其基本思想是通过对象和属性间的关系，构建形式背景，同时定义由对象与属性构成的形式概念，通过概念间的对象包含关系(或者属性间的包含关系) 定义偏序建立概念间的层次关系，称为概念格[8]。随着语义技术的发展，形式概念分析这种作为知识处理和数据分析工具，被广泛应用于本体工程、概念获取等领域。下面就形式概念分析的几个基本定义做简单的概述。

定义1：称三元组（O,A,R）是一个形式背景（Formal Context FC），其中$O=\{O_1,O_2\ldots,O_n\}$是对象集，$A=\{A_1,A_2,\ldots A_m\}$是属性集，R为O和A上的一个二元关系，$R\subseteq O\times A$. 若$(O_i,A_j)\in R$，（$1\le i\le n$，$1\le j\le m$，）则说明对象O_i拥有属性A_j。

对于形式背景FC=(O,A,R)，在对象集$X\subseteq O$和属性集$Y\subseteq A$上分别定义运算：

$$X^*=\{a\in A\mid\forall o\in X,(o,a)\in R\}\quad Y^*=\{o\in O\mid\forall a\in Y,(o,a)\in R\}$$

定义2：设形式背景FC=(O,A,R)，如果一个二元组$(X,Y)(X\subseteq O,Y\subseteq A)$满足$X^*=Y$，且$X=Y^*$，则称(X,Y)是FC=(O,A,R)的一个形式概念，简称概念。其中，X为概念的外延，Y为概念的内涵。

定义3：定义形式背景FC=(O,A,R)的形式概念之间的偏序：$(X_1,Y_1)\le(X_2,Y_2)\Leftrightarrow X_1\subseteq X_2$，或者$Y_2\subseteq Y_1$，则FC=(O,A,R)的所有概念以及概念之间偏序组成的概念层次结构，称为概念格。记为L(O,A,R)。

若两个概念(X_1,Y_1)，(X_2,Y_2)满足偏序关系$(X_1,Y_1)\le(X_2,Y_2)$，则称概念(X_1,Y_1)是下位概念，(X_2,Y_2)是上位概念。

概念格是一种严格的由概念及概念间包含（偏序）关系组成的树状结构，其顶层概念是下层所有概念的概括，上层概念包含下层概念。概念格这一特征类似于概念树，因此使用形式概念分析构建概念树是可行的。

3 单页面内容概念树构建方法

本文提出的单页面内容概念树构建方法是以Web站点中的某一具体页面作为切入点进行分析研究的，其目的是为了更形象化的展示页面的具体分类以及内容结构。概念树构建的基本思想为：首先，获取页面的具体内容信息，并进行预处理操作；其次，以内容中的段落为对象，段落中出现的动、名词为属性，构造形式背景；再次，利用概念格构建方法构造概念格；最后，对概念格中的概念进行剪枝命名转换处理，构建单页面内容概念树。

3.1 页面预处理

本文使用HtmlParser[9]页面解析器对具体的页面进行DOM解析，根据DOM树的结点提取页面的每个段落内容，并按一段落一文件进行存储。利用中科院分

词软件ICTCLAS[10]对每个段落文件进行分词和词性标注，去除停用词，提取其中的动、名词，通过公式（1）对提取的词进行组合，以构造更能表达段落主题的特征词组。

$$Comb(x,y)=\frac{Freq(x,y)}{Freq(x)+freq(y)-Freq(x,y)} \quad (1)$$

其中Freq（x,y）是词x和y在同一段落中同时出现的次数，Freq（x）是词x在段落中出现的次数。

3.2 构建形式背景和概念格

在生成概念格之前，先将经过预处理后的页面，通过生成段落文件以及特征词提取，建立段落与特征词间的二元关系，即行表示为对象、列代表属性的方式形成一个二维表，便可获得相对应的形式背景。形式背景的具体构造过程为：首先，对页面中的每一段落生成的段落文件，在二维表中加入新的一行，第一列对象集中记录段落文件号，并将从该段落文件中提取的动、名词特征项自动添加到形式背景的属性集中，若属性集中已存在特征项，则不重复添加；其次，将二维表中该段落文件号对应的行和属于该段落文件的特征项的列的交叉位置值设为1，或用“X”标示，若该段落文件不具有该特征项，则位置值为空，最后输出形式背景的矩阵表示形式，从而生成初始的形式背景；最后，对所有属性列中出现的同义词进行合并，对形式背景进行简单的简化处理，净化形式背景，降低构造概念格的时间。

将上述生成的形式背景，使用概念格构建工具Conexp1.3[11]来完成概念格的构建。Conexp是德国人Serhiy A. Yevtushenko开发的形式概念分析工具，该工具只需用户提供一个表示形式背景的二维表格文件即可，如表1所示。因此，为了提高概念树的构建效率，本文选择Conexp作为建格工具。

表1 形式背景示例

	教育部	组织	特殊教育改革	试验区	申报	天津市北辰区	创新	教育行政部门
段落1	×	×	×	×	×	×		
段落2		×	×	×			×	×
段落3			×	×			×	
段落4	×	×			×			×

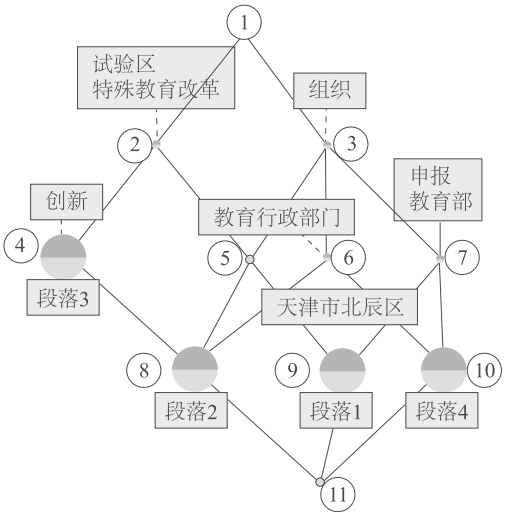

图1 Conexp生成的概念格

3.3 概念格向概念树的转化

由Conexp生成的概念格如图2所示，其中编号1-11是概念格生成后人工添加的，为了方便后续的描述。从图2可以看出概念格类似于树状结构，每一个节点代表一个概念。每个节点对应的属性集是相应概念的内涵，对应的对象集是概念的外延。虽然概念格形似概念树，但两者存在着不同之处，如概念格中某些节点没有对应概念描述等，必须经过修剪命名转换后，才能变成实际意义上的概念树。具体的修剪命名转换规则为：

规则1：概念通常情况下都是由语词来表示，因此剪除所有节点概念的外延，即对象，由属性来表示节点概念，如图1中的段落1、段落2等；

规则2a：删除最底层的节点概念。通过概念格中概念的包含（偏序）关系可知，该概念表达了由所有对象组成的所有属性描述的含义的交集。在实际应用中，不存在此类概念，如图1中的节点11；

规则2b：删除次底层和最底层直接相连且自身不含新属性的概念节点，因为该节点的概念可通过其它节点的概念来描述。如图1的节点8、节点10；

规则3：若一个节点概念只包括一个属性值，则用该属性来命名此概念节点，表达节点概念。如图1中的节点6可用“教育行政部门”来描述概念；

规则4：若一个节点概念包括两个或以上的属性值，则用该节点所含的属性值的交集来命名此概念节点，表达节点概念。如图1中的节点2可命名为“特殊教育改革试验区”概念；

规则5：若一个节点是由几个上位节点合成，即该节

点是多个节点的子节点，且自身没有添加新的属性或对象，则该节点概念可通过合成它的父节点概念的交集来命名，或通过在叙词表或其它知识组织系统中查找这些父概念的共同子概念，并用其命名。如图1中的节点5由节点2和3合成，且没新属性或对象，可用“组织特殊教育改革试验区”来表述此概念；

规则6：若一个节点有多个下位概念节点，且自身没有属性描述，则取其下位节点概念的并集来命名此节点，或通过在叙词表或其它知识组织中查找这些子概念的最低共同祖先概念，并选择一个最合适的概念为其命名。如图1中的节点1可用“教育”来描述概念。

根据上述修剪命名转换规则，图1的概念格转换成的概念树如图2所示。

从该概念树可直观地描述出该网页的大致内容是“关于组织申报教育部特殊教育改革试验区”的，且从顶层概念可知该网页的一级分类应该是“教育”类目。因此，通过构建单页面内容概念树可对网页内容进行重构，并可进一步分析其分类合理性。

4 实验评价

为了进一步验证本文所提出的概念树构建方法，

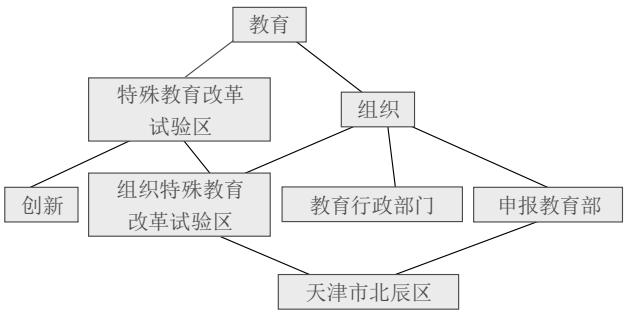

图2 概念格对应的概念树

本文从教育部网站信息公开栏目下[12]的“基础教育”、“职业教育与成人教育”、“高等教育”、“体育卫生与艺术教育”、“教育信息化”、“科学研究”等6个栏目下各分别下载50个页面作为实验对象。使用信息检索中的查准率作为分类准确以及内容重构合理性的评价指标，其计算方法如公式（2）所示。

$$Precision=\frac{|A\cap B|}{|A|}\times 100\% \qquad (2)$$

其中A是待识别分类是否正确的网页集或内容待重构的网页集，B是本文方法分类正确的网页集或内容重构合理的网页集。

使用本文提出的单页面内容概念树构建方法，分别对上述6个栏目的50个页面进行概念树构建，并用概念树的顶层概念来判断网页分类是否正确，概念树直观描述来分析内容重构的合理性，其结果如表2所示。

从表2中可看出，在页面分类方面，“基础教育”、“职业教育与成人教育”、“教育信息化”3个类别方面有较高的准确率，都达到了90%及以上，但在其它3个类别准确率相对低一点，不过也在80%以上。本文从页面具体内容上仔细分析，发现造成这种差别的原因是，在“基础教育”、“职业教育与成人教育”、“教育信息化”类别中的页面基本上都带有了专指度较高的词，如“义务教育”、“职业学校”、“职业教育”、“成人教育”、“网络教学”等；而在“高等教育”、“体育卫生与艺术教育”、“科学研究”栏目下的页面大部分的词都比较宽泛，如“高等学校”、“足球”、“教育经费”等，这些词语分类比较模糊，难以确定其具体类别，特别是“高等教育”栏目的大部分页面的概念树都必须借助叙词表才能凝聚其顶层概念到“高等教

表2 本文方法在识别页面分类和内容重构合理性的实验结果

类别	页面分类			页面分类		
	待分类网页数	正确分类数	Precision	待重构网页数	重构合理的页面数	Precision
基础教育	50	45	90%	50	46	92%
职业教育与成人教育	50	47	94%	50	48	96%
高等教育	50	40	80%	50	45	90%
体育卫生与艺术教育	50	42	84%	50	45	90%
教育信息化	50	46	92%	50	47	94%
科学研究	50	43	86%	50	46	92%

育”；查阅了大量的网页信息，进一步发现页面内容段落较少的网页分类基本上没有成功，这可能是由于内容较少，不适合以段落为对象构建概念树造成的。在内容重构方面，所有类别栏目下的页面重构准确率都达到了90%以上，这说明本文提出的概念树构建方法所生成的单页面内容概念树能较好的展示页面的内容结构，直观地描述页面内容。

5 结语

在海量信息环境下，网络信息多、杂、乱的特点，严重阻碍了用户合理有效地利用网络资源。随着语义技术的发展，Web内容可用性评估逐渐成为人机交互的研究热点。研究发现使用概念树对Web内容进行评估是有效的。因此，本文提出了一种形式概念分析的单页面内容概念树构建方法。本文方法以单个页面作为研究切入点，使用形式概念分析技术在以段落为对象、动名词为属性的背景上进行概念树的构建研究。实验结果表明，本文方法在对网页分类和内容重构方面具有较好的效果，表现出较强的竞争力，但也存在一些不足，如对段落少、内容少的页面不能很好的分类，特别是一句话新闻网页、微博等。在以后的研究工作中可对这一方面进行深入的讨论，考虑以句子为单位来构建概念树。

参考文献

[1] 王战军,瞿斌. 基于概念树的评估指标相关性研究[J]. 系统工程学报,2002,17(6):491-497.

[2] 宣士斌. 基于概念相容性的概念树自动生成算法[J]. 计算机工程与应用,2007,43(6):174- 177.

[3] 毛宇梅. 基于链接结构的Web概念树构建[J]. 计算机工程与应用,2010,46(S):69-71.

[4] 孙亚琳,赵林林,杨小平. 基于主题词表和FCA的网页语义概念树构建研究[J]. 计算机应用研究,2014,31(11):3308-3315.

[5] 杨小平,宇文姝丽,韩佳. 基于概念树的Web语义结构评价[J]. 计算机工程与应用,2011,47 (S2):20-22.

[6] 概念层次树[EB/OL].[2015-04-25]. http://baike.baidu.com/view/2386639.htm.

[7] Ganter B, Wille R. Formal concept analysis, mathematical foundation[M]. Berlin: Springer- Verlag, 1999:68-80.

[8] 杨小平,何伟,孙亚琳,等. TFC-Reducing: 一种基于属性语义距离和规则的文本型形式背景约简方法[J]. 小型微型计算机系统,2012,33(10):2170-2176.

[9] HTMLparser[EB/OL]. [2015-04-25]. http://sourceforge.net/projects/htmlparser/files/.

[10] ICTCLAS[EB/OL]. [2015-04-25] http://baike.baidu.com/link?url=WKc5egl8EouDIPsazWW eM1FoKsOy5dKxHE0ZQIzWyCA1jIU2BAgbk3dSoQlpz33hW1Y8UZr6gi0xfY3KmJv9a.

[11] conexp1.3 [EB/OL]. [2015-04-25].http://sourceforge.net/projects/conexp/.

[12] 教育部信息公开目录[EB/OL]. [2015-04-25]..http://www.moe.gov.cn/publicfiles/business/ htmlfiles/moe/info_category_query/index.html.

作者简介

何伟，男，1978年生，博士，研究方向：本体工程、语义计算，E-mail：hewei@istic.ac.cn。

An Approach for Constructing Content Concept Tree of Single Web Page Based on Formal Concept Analysis

HE Wei[1,2], LI Shuang[2]

(1. Institute of Scientific and Technical Information of China, Beijing 100038, China; 2. Huaihua University, Huaihua 418008, China)

Abstract: In huge information environment, it is serious to affect the effective use of the user with random network information.The Web usability evaluation of page content has become a hot point of research.This paper proposes a method for constructing content concept tree of single web page based on formal concept analysis FCA, which is based on a single page as a starting point,and page paragraph regarded as the object, noun or verb in the paragraph as attributes, using FCA to construct concept lattice, with pruning, named and converted rules to construct the concept tree.Experimental results show that the approach used to build the concept tree has a better effect on page content reconstruction and reasonable classification judgment. It can be used as a method of usability evaluation of Web content.

Keywords: Formal Concept Analysis; Concept Lattice; Constructing Concept Tree

（收稿日期：2015-04-30；编辑：雷雪）