ISSN 1673-9418 CODEN JKYTA8
Journal of Frontiers of Computer Science and Technology
1673-9418/2017/11(01)-0114-10
doi: 10.3778/j.issn.1673-9418.1510016

E-mail: fcst@vip.163.com
http://www.ceaj.org
Tel: +86-10-89056056

针对开源论坛网页的信息抽取研究*

刘春梅[1,2+],郭 岩[1],俞晓明[1],赵 岭[1],刘 悦[1],程学旗[1]
1. 中国科学院 计算技术研究所,北京 100190
2. 中国科学院大学,北京 100190

Information Extraction Research Aimed at Open Source Web Pages*

LIU Chunmei[1,2+], GUO Yan[1], YU Xiaoming[1], ZHAO Ling[1], LIU Yue[1], CHENG Xueqi[1]
1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
2. University of Chinese Academy of Sciences, Beijing 100190, China
+ Corresponding author: E-mail: liucm1005@163.com

LIU Chunmei, GUO Yan, YU Xiaoming, et al. Information extraction research aimed at open source Web pages. Journal of Frontiers of Computer Science and Technology, 2017, 11(1): 114-123.

Abstract: There is a large proportion of forum Web pages generated by open source software. This paper proposes an information extraction method aimed at open source Web pages based on templates. Firstly, a clustering strategy based on the similarity of Web page structure is proposed. The experiment results show that the strategy is superior to the direct classification based on software version. Secondly, a clustering algorithm based on open source software features is proposed. It can cluster large-scale open source forum Web pages based on similarity automatically, and form a marked category. This method not only sharply decreases manual cost on annotation templates, but also increases the accuracy of information extraction.

* The National Basic Research Program of China under Grant Nos. 2012CB316303, 2013CB329602 (国家重点基础研究发展计划(973计划)); the National High Technology Research and Development Program of China under Grant Nos. 2015AA015803, 2014AA015204 (国家高技术研究发展计划(863计划); the National Natural Science Foundation of China under Grant Nos. 61232010, 61173064, 61173008 (国家自然科学基金); the National Key Technology Support Program of China under Grant No. 2012BAH46B04 (国家科技支撑计划); the Technology Innovation and Transformation Program of Shandong Province under Grant No. 2014CGZH1103 (山东省自主创新及成果转化专项); the Medical Imaging Program of Chinese Academy of Sciences under Grant No. KGZD-EW-T03-2 (中科院医学影像项目); the 7th Technology Framework Program of European Union under Grant No. PIRSES-GA-2012-318939 (欧盟第七科技框架计划(FP7)项目).
Received 2015-10, Accepted 2016-01.
CNKI网络优先出版: 2016-01-07, http://www.cnki.net/kcms/detail/11.5602.TP.20160107.1540.006.html

Key words: record locating; Web page clustering; template extraction

摘　要：互联网上大量论坛使用开源软件生成，针对这类论坛，提出了针对论坛网页信息抽取的基于模板的信息抽取方法。首先给出了基于网页结构相似度的簇划分策略，并通过实验证明了该策略优于直接基于软件版本号等直观类别的划分策略；其次提出了基于开源软件特征的聚类算法，能够根据网页相似度将大规模开源软件生成的论坛网页进行有效的自动划分，形成可标注类别。实验表明，该方法不仅保持了基于模板的抽取方法所具有的高准确率的优点，同时弥补了其模板配置与维护代价高的缺点。

关键词：记录定位；网页聚类；模板抽取

文献标志码：A　**中图分类号：**TP181

1　引言

论坛作为一种重要的网络交流平台，承载了丰富的互动信息，论坛网页经过信息抽取后形成的结构化数据对于之后的分析与挖掘都具有重要的应用价值。相对于新闻等其他类型网站，论坛网站的重要特点是：大量论坛网站都基于开源软件构建。在互联网上随机爬取了10万个论坛网页，经统计发现其中开源软件生成的网页的比例为70%左右（见4.1.1节实验）。由于目前针对论坛网页的信息抽取技术在维护代价和准确率等方面尚不能让人满意，本文将基于以上观察，研究针对开源论坛网站的信息抽取方法。

开源论坛网页具有以下特点：

（1）同种开源软件生成的论坛网页有明显的规律特征，这些规律特征包括内容特征和结构特征。内容特征是指网页文本内容的一些指纹信息，如对于discuz生成的网页，在底部会有“Powered By Discuz 2.1”的字样。结构特征是指由相同开源论坛软件生成的网页在DOM树的结构特征方面具有高度相似性。

（2）开源软件生成的论坛网页在结构上相对稳定。因为开源软件只有在出现升级版时，才会出现新类型网页，所以从某种角度上讲，由已知开源软件生成的论坛网页在结构上也是已知且不变的。

论坛网页信息抽取方法通常分为基于模板的抽取方法和全自动抽取方法两类。基于模板的方法具有抽取准确率高的优点，但需要大量人工进行模板配置和维护；全自动的信息抽取方法虽然人工代价小，但准确率相对较低。基于现有方法特点，对开源软件生成的论坛信息进行信息抽取，一种自然的想法是利用开源软件生成网页的规律性，将网页划分成簇，每个簇中的网页具有高度相似性，进而使用基于模板的方法进行高精度抽取；同时，由于簇的数量较少，模板配置和维护代价相对较小。但该方法涉及以怎样的策略对网页进行划分，才能更好地利用规律性的问题；同时，由于论坛信息整体规模较大，人们希望簇划分能够自动进行，这需要找到自动完成簇划分的有效方法。本文将从这两个角度出发提出针对开源论坛软件生成网页的信息抽取方法。

（1）提出基于网页结构相似度的簇划分策略，并通过实验证明该策略优于直接基于软件版本号等直观类别的划分策略。

（2）提出基于开源软件特征的聚类算法，能够根据网页相似度将大规模开源软件生成的论坛网页进行有效的自动划分，形成可标注类别。

使用本文提出的方法抽取开源软件生成的论坛网页，优点在于：具有基于模板抽取方法的高准确率；开源软件生成的网页结构具有很强的规律性，能够使用少量模板覆盖大规模网页，降低模板配置的代价；开源软件生成的网页结构相对稳定，模板失效的可能性很小，从而使得抽取模板的维护代价也很小。从应用的角度看，由于对开源软件论坛页面自动识别较为容易，本文方法不仅可以单独针对开源论坛网站进行抽取，也可以用于提升论坛整体抽取的效果。

本文组织结构如下：第2章介绍相关工作；第3

章介绍核心算法和实现;第4章介绍实验部分;最后进行未来工作的展望。

2 相关工作

本文涉及的相关工作包括两方面:论坛信息抽取和网页聚类。

针对论坛的信息抽取方法主要是全自动信息抽取方法和基于模板的信息抽取方法。全自动论坛信息抽取算法可以分成两类:

(1)单页面重复结构发现。这类方法大多借助网页中记录之间结构的相似性找各条记录所在的节点。例如经典的MDR(mining data records)算法[1]以及其后续改进的算法DEPTA(data extraction based on partial tree alignment)[2]提供了一种多记录节点挖掘法。实际应用中,网页结构有两种特殊情况:论坛页面只包含一楼记录,或者主贴记录与其他记录有不同的结构。由于MDR算法是利用多个相似子树来定位多记录节点,这使得MDR一类的算法只适用于挖掘有多个结构相似的记录节点,而无法处理上述两种情况,导致从整体上降低了论坛信息抽取准确率。本文对此有较好的处理处理方法。

(2)多页面重复结构规约。该类方法以Crescenzi等人的RoadRunner[3]为代表,通过对比多个同类型页面的html串或者DOM树的异同,规约出模板。

近年来,针对上述两类方法只使用了结构信息或文本信息的问题,文献[2,4-5]通过加入Render&Layout过程,引入视觉信息,通过性能牺牲换取了更高的数据记录识别准确率。另外,除了充分挖掘网页通用属性信息,文献[6-7]针对特定的抽取对象,引入了领域知识,并在特定领域的网页内容抽取上取得了很好的成绩。

基于模板的论坛信息抽取方法可以分为两类:包装器归纳方法和模板自动生成方法。而自动化的方法有准确率低的缺点,因此本文采用的抽取方法是包装器归纳法。包装器归纳方法通常由用户在网页源码中标记出需要抽取的信息,然后使用基于规则的机器学习等技术自动归纳出包装器。经典的包装器归纳方法有SoftMealy算法[8]。SoftMealy使用有限状态机来刻画属性之间的转换关系,放松了对属性顺序的要求。因此,SoftMealy支持多对象性、缺值性、多值性和多序性,使得待抽取对象的结构可以存在更多差异,即模板的通用性较好。包装器归纳方法的优点在于一旦生成了包装器,对结构相同或非常相似的网页有很好的抽取效果。但是对于海量且异构的网页来说,包装器的人工维护代价过大。本文使用SoftMealy对开源论坛软件生成网页进行抽取。但在抽取之前,首先使用聚类技术对网页按照结构相似程度划分,然后针对每簇预先配置模板,从而在保持SoftMealy算法本身具有的抽取准确率高,包装器通用性好等优点的同时,大幅度降低SoftMealy包装器的配置与维护代价。

针对于网页聚类,特征选取至关重要。文献[9-10]的聚类特征除了考虑网页结构特征外,还考虑了网页的内容特征。文献[9]的聚类算法的特征是基于网页分块信息的,对每个分块除考虑网页结构特征外,还有锚文本、URL等内容特征;文献[10]的聚类算法的结构特征考虑的是DOM树的分层特征,内容特征包括URL、关键词、锚文本等。另外一些聚类算法如文献[11-14]选取的特征纯粹是基于网页结构的,文献[11]的聚类算法利用局部标签树的编辑距离,而文献[12]的聚类算法利用全局XPath作为网页的特征。本文提出的聚类算法充分利用论坛开源软件生成网页的规律性特征,例如基于记录节点的局部XPath等特征,并选择层次聚类和K中心点算法相融合的H-K算法进行聚类[15]。

3 针对开源软件的信息抽取方法

论坛网页中的帖子列表页面和帖子详情页面包含了丰富且价值很高的信息,本文的研究重点在于解决帖子列表页面和帖子详情页面的抽取问题。如前文所述,开源论坛软件生成的网页有明显的规律特征,这些规律特征包括内容特征和结构特征,而且开源软件生成的论坛网页在结构上相对稳定。考虑到上述规律,将这些开源软件生成的论坛网页划分成多个簇,保证每个簇中的网页具有足够的相似性,从而可以用同一个模板进行抽取。这样做的目的是

在抽取时，为每一个簇配置一个模板即可，既可以获得较高的抽取准确率，又可以大幅度降低模板的配置与维护代价。因此，本章将首先研究簇划分策略，然后提出簇的自动划分方法，最后给出完整的抽取方法。

3.1 簇划分策略

论坛软件本身有软件和版本之分，例如 discuz X2.5，因此最自然的簇划分策略是将同一个软件和版本的开源软件生成的论坛网页划分为一个簇，并为它配置一个模板完成抽取，即按照软件版本对网页进行划分。4.1.2 节对论坛软件版本和同一版本所需抽取模板数量进行了统计分析发现，相同版本号的开源软件生成的网页在结构方面依然存在一定差异，无法使用一个模板完成抽取；同时，进一步分析发现有些不同版本号的开源软件的网页在结构特征方面反而相似度很高。因此利用版本号对网页进行划分的方法难以很好地利用开源软件生成论坛网页间的相似性，不是有效的簇划分策略。

基于模板的抽取方法对网页的 DOM 树特征非常敏感，因此本文根据 DOM 树的结构特征对开源论坛网页做更精确的划分，最终选择了基于 DOM 树结构特征的聚类算法。

4.2 节实验结果表明，和按照软件版本对网页进行划分相比较，按照聚类结果对网页的划分能够保证相同簇中的网页在结构上具有更强的相似性，使得每一簇网页可以共用一个信息抽取模板，而且只需对簇的中心页面配置模板即可。由于开源论坛软件生成的网页结构相对稳定，使得人们预先得到的聚类结果，以及对每一个簇配置的抽取模板都相对稳定。这些特点都大幅度降低了生产环境中对抽取模板的失效检测与维护的代价。

3.2 基于聚类的划分方法

为了描述本文提出的基于聚类的划分方法，给出以下定义：

（1）记录节点和记录统领节点。网页的 DOM 树中，每条数据记录中承载的子树的根节点称为记录节点，在图 1 中是 div[class,bm_c]以及它的右兄弟 div 节点。记录统领节点就是所有记录节点的父节点，在图 1 中是 div[class,vt]节点。

（2）主贴节点。主贴记录也是一种特殊的数据记录，在网页中，它与其他数据记录在格式上有很大的不同，为了区别其他数据记录，将主贴记录子树的根节点成为主贴节点。

（3）子树特征。子树特征是从该子树的根节点到所有结束节点的 XPath 特征的集合。

每个 XPath 特征是一个字符串，由途经的标签名称和 class 属性构成。如果 class 属性中有数字，则一律转换成\d+的正则表达式形式。如图 1 中 DIV [class,bm_c]节点的特征集合为：

{div[class,bm_user]|, div[class,bm_user]|a, div[class, bm_ user]|em, div[class,mes]|div[class,postmessage]}

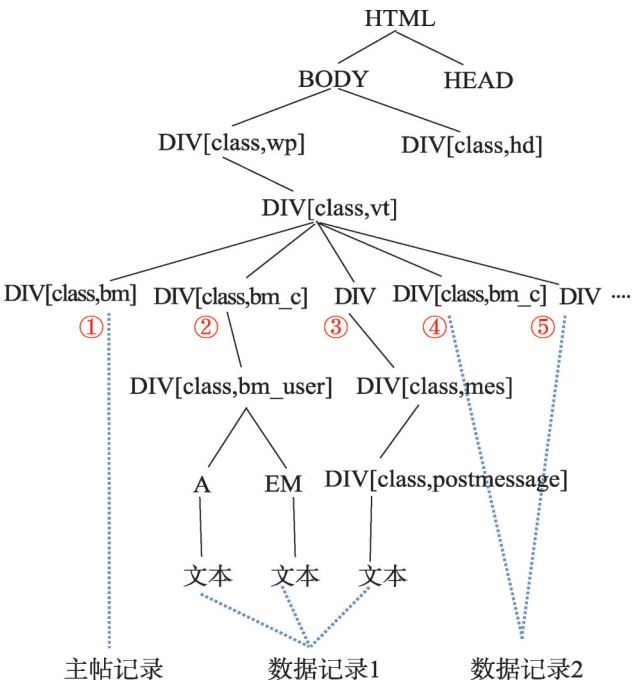


Fig.1 DOM tree

图 1 DOM 树举例

根据特征的功能，本文将记录节点的特征分为两类：

（1）聚类特征。网页中每个记录节点下的子树特征是记录节点的聚类特征。

（2）定位特征。记录节点所在的位置特征，以及记录节点的格式特征统称为记录节点的定位特征，进而提取聚类特征。

本文提出的基于聚类的划分方法需要解决的关键问题如下：

(1)定位特征库的构建。在进行聚类之前，为了减少网页中导航、广告等造成的噪音特征，首先需要利用定位特征定位到网页中所有的记录节点。为了能够增量地利用这些具有共性的定位特征，使用大量开源软件生成的网页进行离线训练，利用相似子树聚类等技术自动学习出记录节点的定位特征，并将其存储到定位特征库中。

(2)结构相似网页的聚类。在聚类过程中，首先利用定位特征库中的特征找到网页中所有的记录节点，然后在每个记录节点下抽取局部XPath作为聚类特征进行聚类，最后为每个聚类的中心页面配置模板。为了得到聚类的中心页面，本文采用的聚类算法是基于层次聚类和K中心聚类的融合算法H-K算法。

3.2.1 定位特征库的构建

通过分析，将记录节点和主贴节点分别在论坛列表页面和帖子页面中的位置情况分为图2的4种类型。

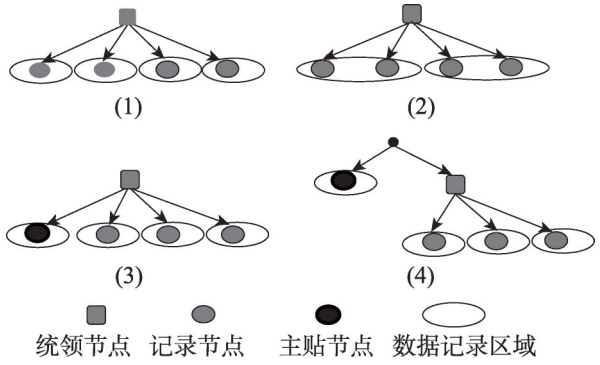


Fig.2 Four cases of page structures

图2 论坛网页结构的4种情况

对于前两种情况，需要查找到重复子树集合，集合中每个重复子树的根节点即记录节点。后两种情况分别表示含主贴节点的情况。基于以上观察，构建定位特征库主要有以下3步：

步骤1 记录节点定位

后序遍历网页的DOM树，保证在遍历到某个节点时，其孩子节点的特征已经提取完毕。在遍历过程中，对当前遍历的节点做以下3个操作：

(1)判断候选记录统领节点

因为记录统领节点具有这样的特性：它的重复子树数目大于1，所以需要进行重复子树的查找，利用聚类的方法，将重复子树聚到一个簇中。有的簇中重复子树个数大于1，则认为当前遍历节点有可能是记录统领节点，记为候选统领节点。

(2)判断候选记录节点

一个数据记录的子树可能存在于多个簇中，如图2的第二种情况所示，因此需要将属于一条数据记录的簇进行合并，合并的条件是：一个簇中按顺序存储着每个子树的位置索引，如果簇的子树数目相等，且簇间子树的位置索引间隔依次相等，则可将这些簇进行合并。最后得到的所有重复子树个数大于1的簇，这些簇中的子树节点都是候选记录节点。

(3)筛选记录节点

为了保证记录节点定位的准确性，需要使用一些约束条件从候选记录节点中筛选出真正的记录节点。约束条件有如下5点：

① 一般记录统领节点和记录节点的标签为div、table。

② 每一条数据记录中都需要有时间序列，如2014-12-26 12:14，前天14:24。

③ 如果记录节点含有Id的属性，那么它的Id为数字且长度大于3。

④ 记录统领节点的深度一般不超过11。

⑤ 满足上述条件的节点中，深度最大的节点作为统领节点。

对于单记录页面或者噪声页面，使用以上约束条件筛选后，将不会抽出任何记录节点。

步骤2 主贴节点的发现

研究发现主贴节点具有以下特征：

(1)主贴节点是第一个记录节点的左兄弟或者统领节点的左兄弟。

(2)主贴节点中含有时间序列信息且文本内容满足一定的长度。

因此，在找到记录节点后，根据以上特征继续发现主贴节点。

步骤3 定位特征存储

找到记录节点和主贴节点后，需要从中提取定位特征。提取过程如下：

（1）因为标签节点的ID属性在整个网页中有唯一性，所以如果最后查找的记录节点有ID的属性，则将该ID换成正则表达式形式并作为定位特征存储，如post_123456的ID转换成post_\d+形式。

（2）如果不含有ID属性，则将该节点以及其3个前驱节点的XPath特征作为记录节点的定位特征存储。

因为每个定位特征都是一个字符串，所以在使用定位特征库进行记录节点的定位时，可以将定位特征库以哈希表的形式存在内存中，从而加快检索效率。

3.2.2 网页聚类

在聚类过程中，首先遍历DOM树，查找所有满足定位特征的节点，然后提取出该节点下的子树特征作为网页的聚类特征。

本文的聚类算法采用层次聚类和K中心点融合的H-K算法。K中心算法的初始K值和初始聚类中心的选择由基于阈值的层次聚类得到。经过多次实验，发现当阈值设为2时可以很好地划分结构相似的网页。将距离小于阈值的网页放到一个初始簇中。

网页i和网页j间的距离定义如下：

$$dist_{i,j}=\frac{len_i\times len_j+1}{overlap+1}-1$$

其中，len_i表示网页i中特征的个数；len_j表示网页j中特征的个数；$overlap$表示两个网页相同的特征的个数。

得到初始簇分布后，利用K中心算法进行迭代，当簇分布在两次迭代中完全一样时，迭代停止，得到最终的簇分布结果。

3.3 开源论坛网页的信息抽取方法

在实际应用中，针对每个簇的中心页面，使用SoftMealy算法配置模板。信息抽取的流程如图3所示。其中虚线框中的过程即是上文提到的两个技术关键点。

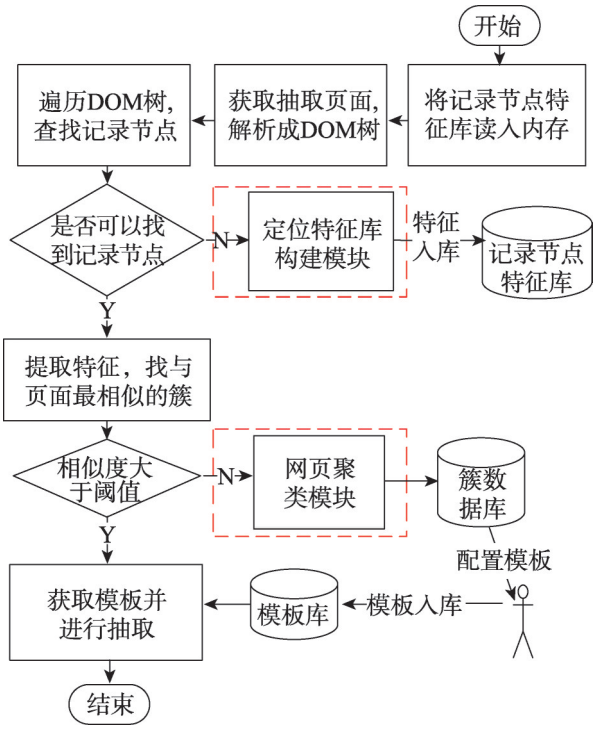


Fig.3 Extraction flow

图3 抽取流程图

4 实验

实验数据集是作者利用爬虫在互联网上随机爬取的10万个论坛网页，来源于2 455个网站。其中帖子列表页面有40 000个，帖子详情页面有60 000个。

4.1 开源软件网页统计

4.1.1 开源论坛网页的比例

本文利用开源论坛网页的内容特征，对数据集中的开源论坛网页所占比例以及各个开源软件比例进行统计，得到如图4的统计结果。

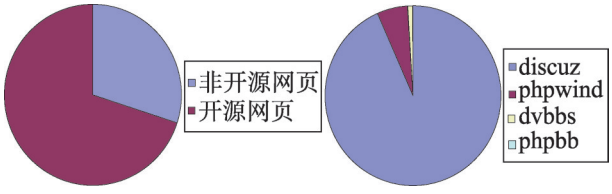


Fig.4 Proportion of open source Web pages and software

图4 开源论坛网页以及软件的比例

由图4可见，数据集中的开源论坛网页所占比例为70%，约7万个。这些网页也是下面实验主要的训练集。

4.1.2 开源软件的版本

本文利用开源软件生成的网页的内容特征，获知每个网页在开源软件中的具体版本。对4.1.1节中统计得到的4个常见的开源软件，继续统计每个开源软件有多少种版本，得到的统计结果如表1所示。

Table 1 Versions of open source softwares

表1 开源软件版本统计

开源软件	discuz	phpwind	dvbbs	phpbb
版本数目	23	11	5	1

4.1.3 相同版本的模板数目

相同版本号的网页的结构应该是有一些共性的，为了验证相同版本号的网页是否可以用一个模板进行配置，本文选择了比例最大的discuz软件中一些常见的版本号，对于每个版本号在数据集中随机选择了200个网页，进行模板数目的统计，得到如表2所示的结果。

Table 2 Versions and templates statistics

表2 版本号与模板数量统计

版本号	7.0.0	7.2.0	X1.5	X2.0
模板数	2	2	2	6
版本号	X2.5	X3.0	X3.1	X3.2
模板数	8	6	11	15

以上实验也表明了相同版本号的网页并不能全部用一个模板进行抽取。

4.2 记录节点定位

因为记录节点定位需要人工判断正确性，所以本部分实验在数据集中随机选择200个开源论坛网页，这200个网页来自200个不同的网站，其中版面页面以及帖子详情页面各占一半。

4.2.1 评价指标

根据抽取的结果，可将抽取结果分为5种情况：

A，多记录页面，抽取出正确的记录节点；

B，多记录页面，抽取出错误的记录节点；

C，多记录页面，没抽取出记录节点；

D，对于单记录页面，没有抽取出结果；

E，对于单记录页面，抽取出结果。

可以看出，如果A+D所占的比例越大，实验效果越好。

4.2.2 实验结果及分析

本部分实验主要是与文献[1]提到的多记录挖掘算法MDR进行对比。比较结果表3所示。

Table 3 Record mining algorithms comparison

表3 记录挖掘算法比较

算法	A数目	B数目	C数目
本文算法	142	1	2
MD算法	127	8	10
算法	D数目	E数目	A+D比例
本文算法	55	0	98.5%
MD算法	0	55	63.5%

本文算法旨在降低B和E的两种情况。因为这两种情况可能会产生比较坏的后果，会造成对网页记录产生错误定位，从而提取出错误的特征；而对于C这种情况，在训练数据量大的情况下，一个网页的数据记录由于一些原因无法抽取，却可以通过别的网页来抽取，最终影响较小。

最终特征库生成的训练集是数据集中7万个开源论坛网页，共生成了77条记录特征。

4.3 网页聚类

为验证3.2.2节网页聚类的效果，本部分实验利用之前生成的定位特征库，随机选择200个开源论坛网页作为训练集。聚类结果以是否能用一个模板抽取为评价标准，人工构建了标准答案集。

4.3.1 评价指标

利用准确率、召回率以及$F1$值对聚类结果进行评价。设整个网页数目为N，则两两构成一对，则该数据集中的网页对为$N(N-1)/2$个。网页聚类结果有如下4种情况：

TP，同一类的网页被分到同一个簇；

TN，不同类的网页被分到不同簇；

FP，不同类的网页被分到同一个簇；

FN，同一类的网页被分到不同簇。

准确率为 $P=\frac{TP}{TP+FP}$，召回率为 $R=\frac{TP}{TP+FN}$，$F1$值为 $F_1=\frac{2PR}{P+R}$。

4.3.2 实验结果及分析

选择聚类算法FPC[10]进行对比实验。FPC聚类

算法的特征选取了DOM树的分层信息作为聚类特征。结果如表4所示。

Table 4 Comparison of cluster algorithms

表4 聚类算法比较

算法	准确率/%	召回率/%	*F*1值/%
FPC算法	33.6	98.2	50.1
本文算法	91.0	97.2	94.0

因为最终需要按照是否能共用一个模板来划分簇，所以对每一个网页的结构非常敏感。通过表2可以看到，对于论坛网页的聚类，FPC算法效果远不如本文算法。原因如下：一是FPC算法采用的是*K*Means算法，*K*Means算法对噪声比较敏感，聚类效果不如本文改进的H-*K*算法；二是FPC算法采用的特征是DOM树的分层特征，相比之下，本文算法的附带class属性的XPath特征对DOM树的结构有更好的表示。

Table 5 Extraction results

表5 抽取测试结果

论坛网站名称	页面类型	成功抽取/总共测试
朝九晚五广州论坛	帖子详情页面	10/10
	帖子列表页面	10/10
韩剧社区	帖子详情页面	10/10
	帖子列表页面	10/10
PLU论坛	帖子详情页面	10/10
	帖子列表页面	10/10
声同论坛	帖子详情页面	10/10
	帖子列表页面	10/10
PCBETA论坛	帖子详情页面	10/10
	帖子列表页面	10/10
一元一论坛	帖子详情页面	10/10
	帖子列表页面	10/10
茶韵墨香论坛	帖子详情页面	10/10
	帖子列表页面	10/10
中国球迷论坛	帖子详情页面	10/10
	帖子列表页面	10/10
绍兴会计网	帖子详情页面	10/10
	帖子列表页面	10/10
中国服装人论坛	帖子详情页面	10/10
	帖子列表页面	10/10
HASEEBBS论坛	帖子详情页面	10/10
	帖子列表页面	10/10
落伍者论坛	帖子详情页面	10/10
	帖子列表页面	10/10
帝国论坛	帖子详情页面	10/10
	帖子列表页面	10/10
东方网摄影论坛	帖子详情页面	10/10
	帖子列表页面	10/10

4.4 抽取测试

最终的聚类和标注的训练集是数据集中7万个开源论坛网页，测试数据来源于凤凰网发布的论坛排行榜中的14个开源论坛的网站，这14个网站没有在训练集中出现过。最终共配置了60个模板。对于每一个网站各自抽取了10个帖子详情页面以及10个帖子列表页面，测试结果抽取准确率可达到100%，具体结果如表5所示。

5 小结

本文针对开源论坛网页规律性强，所占比重大的特点，先利用大量训练网页构建定位特征库，然后利用定位特征定位记录节点，获取节点下的XPath特征作为聚类特征，最后将这些网页进行聚类，为每一个聚类的中心页面人工配置模板。实验表明，这些模板很大程度上覆盖了开源软件生成的网页，在降低人工标注代价的前提下，大幅度提高了抽取的准确率。

References:

[1] Liu Bing, Grossman R, Zhai Yanhong. Mining data records in Web pages[C]//Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, Aug 24-27, 2003. New York: ACM, 2003: 601-606.

[2] Zhai Yanhong, Liu Bing. Web data extraction based on partial tree alignment[C]//Proceedings of the 14th International Conference on World Wide Web, Chiba, Japan, May 10-14, 2005. New York: ACM, 2005: 76-85.

[3] Crescenzi V, Mecca G, Merialdo P. RoadRunner: towards automatic data extraction from large Web sites[C]//Proceedings of the 27th International Conference on Very Large Data Bases, Roma, Italy, Sep 11-14, 2001. San Francisco, USA: Morgan Kaufmann Publishers Inc, 2001: 109-118.

[4] Grigalis T. Towards Web-scale structured Web data extrac-

tion[C]//Proceedings of the 6th ACM International Conference on Web Search and Data Mining, Rome, Italy, Feb 6-8, 2013. New York: ACM, 2013: 753-758.

[5] Huang Wuguan, Zhu Ming, Yin Wenke. Web information automatic extraction based on DOM tree and visual feature[J]. Computer Engineering, 2013, 39(10): 309-312.

[6] Alvarez M, Pan A, Raposo J, et al. Extracting lists of data records from semi-structured Web pages[J]. Data & Knowledge Engineering, 2008, 64(2): 491-509.

[7] Yang Zhou, Zhuo Lin, Zhao Pengpeng, et al. Automatic extraction method for product data records[J]. Computer Engineering, 2010, 36(23): 262-265.

[8] Hsu C-N, Dung M-T. Generating finite-state transducers for semi-structured data extraction from the web[J]. Information Systems, 1998, 23(8): 521-538.

[9] Fan Yixing, Guo Yan, Li Xipeng, et al. A multi-level page clustering method based on page segmentation[J]. Journal of Shandong University: Natural Science, 2015, 50(7): 1-8.

[10] Yu Jun, Guo Yan, Zhang Kai, et al. FPC: fast increment clustering for large scale Web pages[J]. Journal of Chinese Information Processing, 2016, 30(2): 182-188.

[11] Liu Yunfeng. A text information extraction algorithm based on tag XPath clustering[J]. Computer Applications and Software, 2010, 27(11): 199-202.

[12] Li Rui, Zeng Junyu, Zhou Siwang. Improved Web page clustering algorithm based on partial tag tree matching[J]. Journal of Computer Applications, 2010, 30(3): 818-820.

[13] Han Pu, Wang Ze. Information extraction for Web forum based on repeated pattern[J]. Journal of Nanjing Normal University: Engineering and Technology Edition, 2010, 10(3): 74-77.

[14] Chen Ting, Liu Jiayong, Xia Tian, et al. Information extraction research based on panel-structured Web BBS[J]. Journal of Chengdu University of Information Technology, 2009, 24(1): 1-4.

[15] Zhang Hongyun, Li Pingping. A *K*-means clustering algorithm based on hierarchy[J]. Software Time, 2010, 26(4): 228-229.

附中文参考文献：

[5] 黄武冠, 朱明, 尹文科. 基于DOM树和视觉特征的网页信息自动抽取[J]. 计算机工程,2013, 39(10): 309-312.

[7] 杨舟, 卓林, 赵朋朋, 等. 一种针对商品数据记录的自动抽取方法[J]. 计算机工程, 2010, 36(23): 262-265.

[9] 范意兴, 郭岩, 李希鹏, 等. 一种基于网页块特征的多级网页聚类方法[J]. 山东大学学报: 理学版, 2015, 50(7): 1-8.

[10] 余钧, 郭岩, 张凯, 等. FPC: 大规模网页的快速增量聚类[J]. 中文信息学报, 2016, 30(2): 182-188.

[11] 刘云峰. 一种基于标签路径聚类的文本信息抽取算法[J]. 计算机应用与软件, 2010, 27(11): 199-202.

[12] 李睿, 曾俊瑀, 周四望. 基于局部标签树匹配的改进网页聚类算法[J]. 计算机应用, 2010, 30(3): 818-820.

[13] 韩普, 王泽. 基于重复模式的论坛信息抽取研究[J]. 南京师范大学学报: 工程技术版, 2010, 10(3): 74-77.

[14] 陈挺, 刘嘉勇, 夏天, 等. 基于平板型Web论坛的信息抽取研究[J]. 成都信息工程学院学报, 2009, 24(1): 1-4.

[15] 张红云, 李萍萍. 一种基于层次聚类的*K*均值算法研究[J]. 软件时空, 2010, 26(4): 228-229.

LIU Chunmei was born in 1991. She is an M.S. candidate at Institute of Computing Technology, Chinese Academy of Sciences. Her research interests include information extraction and data mining, etc.

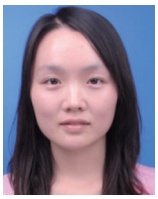
刘春梅(1991—),女,河北邯郸人,中国科学院计算技术研究所硕士研究生,主要研究领域为信息抽取,网页数据挖掘等。

GUO Yan was born in 1974. She received the Ph.D. degree in network information processing from Institute of computing technology, Chinese Academy of Sciences in 2004. Now she is an associate researcher at Institute of Computing Technology, Chinese Academy of Sciences. Her research interest is Web information processing.

郭岩(1974—),女,陕西西安人,2004年于中国科学院计算技术研究所获得博士学位,现为中国科学院计算技术研究所高级工程师,主要研究领域为网络信息处理。

YU Xiaoming was born in 1977. He received the Ph.D. degree in computer system architecture from University of Chinese Academy of Sciences in 2008. Now he is an associate professor at Institute of Computing Technology, Chinese Academy of Sciences. His research interests include Internet search and mining, etc.

俞晓明(1977—)，男，山东青岛人，2008年于中国科学院大学获得博士学位，现为中国科学院计算技术研究所高级工程师，主要研究领域为互联网搜索与挖掘等。



ZHAO Ling was born in 1987. She received the M.S. degree in human-computer interaction from Institute of Software, Chinese Academy of Sciences in 2012. Now she is an engineer at Key Laboratory of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences. Her research interests include Web crawling and structured data extraction, etc.

赵岭(1987—)，女，河南南阳人，2012年于中国科学院软件研究所人机交互与智能信息处理实验室获取硕士学位，现为中国科学院计算技术研究所网络数据科学与技术重点实验室工程师，主要研究领域为Web信息采集，结构化文本抽取等。

LIU Yue was born in 1971. She is an associate professor at Institute of Computing Technology, Chinese Academy of Sciences. Her research interests include information retrieval, data mining, complex network analysis and social computing, etc.

刘悦(1971—)，女，博士，中国科学院计算技术研究所副研究员，主要研究领域为信息检索，数据挖掘，复杂网络分析，社会计算等。

CHENG Xueqi was born in 1971. He received the Ph.D. degree in computer science from Institute of Computing Technology, Chinese Academy of Sciences in 2006. Now he is a professor and Ph.D. supervisor at Institute of Computing Technology, Chinese Academy of Sciences. His research interests include Web information retrieval, social media analytics and network data science, etc.

程学旗(1971—)，男，安徽安庆人，2006年于中国科学院计算技术研究所获得博士学位，现为中国科学院计算技术研究所研究员、博士生导师，中国科学院网络数据科学与技术重点实验室主任，主要研究领域为网络信息检索，社会媒体分析，网络数据科学等。发表学术论文100余篇，主持10余项国家自然科学基金、973计划、863计划等重要科研项目，2014年获国家杰出青年科学基金资助。