文章编号：1007-5321(2017)增-0085-04 DOI:10.13190/j.jbupt.2017.s.019

HTML 页面中的文献记录分析算法

曾庆涛[1,2]， 解 凯[1]， 李业丽[1]， 王欣刚[3]， 叶宇姗[1]， 马少平[2]

(1.北京印刷学院 信息工程学院，北京 102600；2. 清华大学 计算机科学与技术博士后流动站，北京 100084；
3. 国家新闻出版广电总局 广播电视卫星直播管理中心，北京 100045)

摘要：为了使出版机构能够及时从大量网页中发现所需文献，需要设计能够从超文本标记语言页面中自动提取文献信息的算法.为此，设计了基于条件随机场的文献记录分析算法：首先，设计了文档对象树的分割算法，通过分割标记将页面数据分成独立的部分，这些数据块由标签和文本序列构成；随后，将该序列作为条件随机场模型的特征向量，建立文献信息标记模型；最后，设计启发式算法，从标记模型中提取文献信息数据，并通过实验验证了其有效性.

关 键 词：数字出版；条件随机场；文献记录分析

中图分类号：TP393 **文献标志码**：A

Analysis Algorithm of Reference Record in HTML Page

ZENG Qing-tao[1,2]， XIE Kai[1]， LI Ye-li[1]， WANG Xin-gang[3]， YE Yu-shan[1]， MA Shao-ping[2]

(1. School of Information Engineering, Beijing Institute of Graphic Communication, Beijing 102600, China;
2. Postdoctoral Research Station in Computer Science and Technology, Tsinghua University, Beijing 100084, China;
3. Broadcast and Television Direct Broadcasting Satellite Management Center, The State Administration of Press, Publication, Radio, Film and Television, Beijing 100045, China)

Abstract: With rapid development of Internet, web pages have become the main sources of information. In order to make publishing agencies timely find necessary references from large number of pages, it is necessary to design a reference information extraction algorithm to get useful references information from hyper text markup language pages. A reference analysis algorithm based on conditional random fields was proposed. Firstly, a document object tree segmentation algorithm was designed. Through classifier the web page data were divided into separate parts, and these data blocks were composed of tags and text sequences. Subsequently, these sequences were taken as characteristic vectors of conditional random field model to establish reference information labeling model. Finally, a heuristic algorithm was presented to extract reference information data from the labeling model, and validity of this algorithm was verified by experiments.

Key words: digital publishing; conditional random field; reference analysis

随着互联网的快速发展，出版机构越来越依赖从网络中搜集作者和出版内容信息的能力.而文献每年都会数以万计地增长，面对如此多的文献数据，如何从出版物网页中快速、准确地提取文

收稿日期：2016-05-26

基金项目：北京市教委科技创新服务能力建设项目(PXM2016_014223_000025)；北京印刷学院校级重点项目(ea201507)；北京印刷学院教师队伍建设—博士启动金项目(27170116005/062)；北京印刷学院科研项目—出版物数据资产评估实验室建设项目(20190116005/006).

作者简介：曾庆涛(1982—)，男，讲师，E-mail：jiakechongbeijing@163.com.

献信息成为研究热点[1-2]. 与此同时,伴随互联网+的快速普及和迅速发展,各类学术作者的相关信息可以在网络中取得. 如何从大量设计样式灵活多样的作者个人信息页面中提取文献信息成为一个难点[3-8].

针对上述问题,设计了基于条件随机场的文献记录分析算法(CRFRA, conditional random fields based reference analysis-algorithm). 首先,建立分割标记集合,在此基础上设计了文档对象模型(DOM, document object model)树分割算法,并对输入的DOM树进行分割;随后,基于条件随机场建立已分割数据块的文献信息标记模型;最后,设计了文献信息提取的启发式算法,利用该算法提取文献相关信息,并将其存储在数据库中. 最后,通过实验对CRFRA算法和条件随机场(CRF, conditional random fields)算法进行了对比,验证了CRFRA算法的有效性.

1 问题模型

1.1 问题描述

文献信息是描述文献作者、题目等重要信息的半结构化字符串,经常与大量其他信息一同出现在出版页面之中,这就增大了提取文献信息的难度. 此外,文献元数据转换为半结构化字符串的样式灵活多样,即字符串在超文本标记语言(HTML, hyper text markup language)中的描述方法是灵活多样的. 网页中文献信息的样式依赖于网页设计者的个人喜好,在同一页面上的2条字符串可能有不同风格的HTML结构,增大了分析和提取文献信息的难度.

1.2 问题处理流程

针对文献信息提取遇到的问题,设计了基于条件随机场的文献记录分析算法,主要包括:DOM树分割算法,文献信息标记建模和文献信息提取算法. 首先,利用分割标记对输入的DOM树进行分割;随后,基于条件随机场建立已分割数据块的文献信息标记模型;最后,利用文献信息提取算法,提取相关信息,并将其存储在数据库中. 具体处理流程如图1所示.

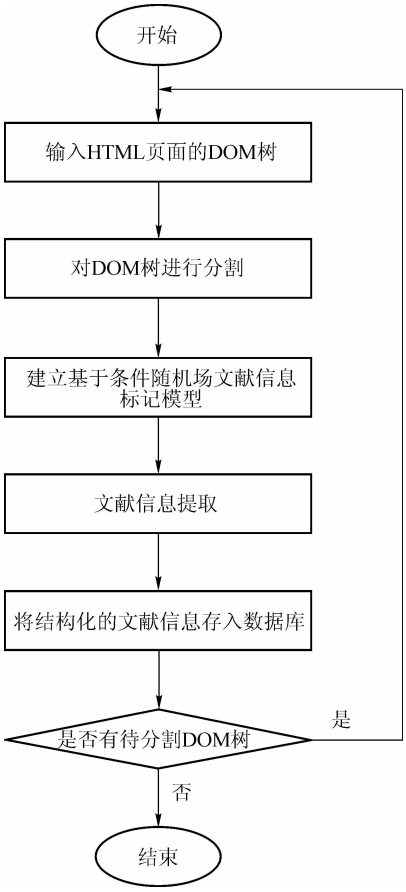

图1 问题处理流程

2 文献记录分析算法

2.1 分割算法

主要将 < tile >, < DIV >, < P >, < LI >, < TR >, < meta name >,作为分割标记,对DOM树 T 进行分割. 分割算法思路为:首先,遍历DOM树 T 中没访问过的节点 t_i,如果 t_i 含有分割标记,并且其子树中不含分割标记,那么将 t_i 存入数据块分割集 B,标记 t_i 为已访问;随后,处理 t_i 的子树中含有分割标记的情况,并标记 t_i 为已访问,继续访问 T 中未被访问过的节点;最后,已遍历 T 的全部节点后,得到 T 的数据块分割集 B.

具体算法步骤如下:

步骤1 输入DOM树 T 和分割标记集合 M;

步骤2 遍历 T 的未访问节点 $t_i \in T$,转步骤3;

步骤3 如果 $t_i \in M$,且其子树 $t_i' \notin M$,则转步骤4;否则,标记 t_i 为已访问,转步骤2;

步骤4 将 t_i 加入数据块分割集 B,标记 t_i 为已访问,转步骤5;

步骤 5　如果已经遍历 T，转步骤 6；否则，转步骤 2；

步骤 6　分割算法结束，B 为 T 的数据块集合.

2.2　基于条件随机场的文献信息标记模型

令 $\boldsymbol{x}=\{x_1,x_2,\cdots,x_n\}$ 为文本的观察序列，$\boldsymbol{y}=\{y_1,y_2,\cdots,y_n\}$ 是与文本观察序列关联的标记序列. 观察序列和标记序列的条件（随机场 CRF，conditional random fields）模型：

$$p(\boldsymbol{y}\mid\boldsymbol{x},\lambda)=\frac{1}{z(\boldsymbol{x})}\exp\Big(\sum_{i=1}^{n}\sum_{j}\lambda_j f_j(y_{i-1},y_i,\boldsymbol{x},i)\Big) \tag{1}$$

$$z(\boldsymbol{x})=\sum_{j}\exp\Big(\sum_{i=1}^{n}\sum_{j}\lambda_j f_j(y_{i-1},y_i,\boldsymbol{x},i)\Big) \tag{2}$$

其中：$f_j(y_{i-1},y_i,\boldsymbol{x},i)$ 是特征函数，是状态特征函数和转移特征函数的统一形式表示，特征函数是二值函数，其值域为 $f_j\in\{0,1\}$；$\boldsymbol{\lambda}=(\lambda_1,\lambda_2,\cdots,\lambda_n)$ 是特征函数对应的特征权重；$z(\boldsymbol{x})$ 是归一化因子.

$$f(y_{i-1},y_i,\boldsymbol{x},i)=\begin{cases}g(\boldsymbol{x},i) & y_{i-1}\rightarrow\text{next}=y_i\\ 0, & \text{其他}\end{cases} \tag{3}$$

$$g(\boldsymbol{x},i)=\begin{cases}1, & x_i\in\text{lexicon}[y_i]\\ 0, & \text{其他}\end{cases} \tag{4}$$

这里通过实例说明上式的含义：当 y_{i-1} 标记为“出版物类型”，y_i 标记为“保存地点”，并且 x_i 是地名集合中的元数据时，特征函数 $f(y_{i-1},y_i,\boldsymbol{x},i)=1$；否则，$f(y_{i-1},y_i,\boldsymbol{x},i)=0$.

2.3　文献信息提取算法

通过 CRF 模型标记包含文献信息的数据块 $R_j=\{r_1,r_2,\cdots,r_{m_j}\}$，随后从这些数据块中提取出文献信息. 在这个过程中可以采用 Viterbi 算法，但是，当该算法的观察序列较长时计算量非常大，这种穷举遍历的搜索算法降低了搜索效率. 为此，设计了启发式算法：首先，在数据块中搜索 <tile>，<DIV>，<P>，<LI>，<TR>，<meta name>，以确定文献信息识别的起始位置；随后，利用式（5）和式（6）对数据块进行筛选；最后，将 $g_F(r_j,h)=1$ 和 $g_B(r_k,h)=1$ 对应的数据块中包含的文献信息提取出来，存入数据库.

$$g_{\mathrm{F}}(r_j,h)=\begin{cases}1, & \{r_j,r_{j+1},r_{j+2}\}\in h,\ 1\leqslant j\leqslant k-2\\ 0, & \text{其他}\end{cases} \tag{5}$$

$$g_{\mathrm{B}}(r_k,h)=\begin{cases}1, & \{r_{k-2},r_{k-1},r_k\}\in h,\ j\leqslant k\leqslant m_j\\ 0, & \text{其他}\end{cases} \tag{6}$$

其中：$h=\{$作者姓名列表，文献题目，出版物类型$\}$，r_j、r_{j+1}、r_{j+2}、r_{k-2}、r_{k-1}、r_k 为数据块中的标记序列.

3　实验分析

实验分别以 3 个中文数字图书馆中的 200 个论文出版页面和 200 位作者的个人信息页面为基础，分别对 CRFRA 算法和 CRF 算法进行对比分析，其对比结果如图 2 所示. 从图 2 中不难看出，在分别采用 CRFRA 和 CRF 算法对 3 个中文数字图书馆的论文出版页面进行文献信息提取时，CRFRA 算法的准确率能够保持在 80% 左右，而 CRF 算法准确率也保持在 70% 以上. 但是，当采用这两种算法对作者个人主页中的文献信息进行提取的时候，CRFRA 算法和 CRF 算法的准确率均有大幅下降. 产生这种现象的主要原因是由于数字图书馆的论文出版页面设计较为规范，而个人主页的设计较为多样，且页面中含有大量其他信息，进一步增大了文献提取的难度.

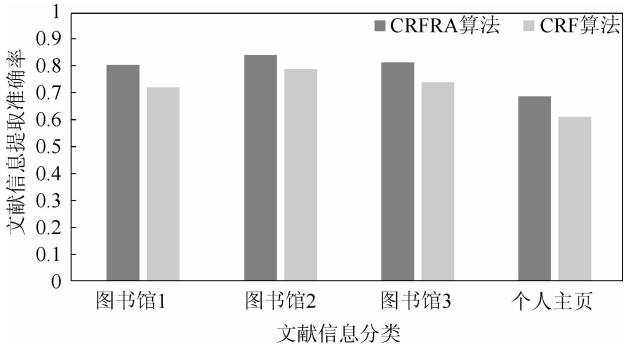

图 2　文献信息提取准确率对比

4　结束语

针对由文献出版页面及作者信息页面设计样式多样造成的文献信息提取困难的问题，设计了基于 CRF 的文献记录分析算法. 对文献信息提取问题的处理流程进行了细化，梳理其中的关键问题，并设计了相应的解决方法. 实验结果表明 CRFRA 算法能够提升文献信息提取的准确率，该算法的有效性得到了验证.

参考文献：

[1] 湛江. 文献检索统计中易被漏检和错误归类的高校学报[J]. 中国科技期刊研究, 2015, 26(9): 1005-1008.
Zhan Jiang. The journals of universities easily missed or wrongly classified in statistical analysis[J]. Chinese Journal of Scientific and Technical Periodicals, 2015, 26(9): 1005-1008.

[2] 孙颖,崔洁爽,陈扬.关键词共现分析技术在图书馆文献检索中的应用——以心理学为我国“五位一体”战略布局服务为例[J]. 图书馆工作与研究, 2015(11): 45-49.
Sun Ying, Cui Jieshuang, Chen Yang. Keywords co-occurrence analysis technology in the library literature retrieval application—to psychology for China “one of five” strategic layout of the service as an example[J]. Library Work and Study, 2015(11): 45-49.

[3] 林岚.认知弹性理论在文献检索教学中的应用[J]. 图书馆, 2010(2):119-120.
Lin Lan. Application of cognitive flexibility theory on document retrieval teaching[J]. Library, 2010(2): 119-120.

[4] 张莉.文献检索方式的发展与提高期刊影响力[J]. 编辑学报, 2005, 17(2): 124-125.
Zhang Li. Evolution of literature retrieval and improvement of the journal's influence[J]. Acta Editologica, 2005, 17(2): 124-125.

[5] 张佳, 窦丽华, 陈杰. 科技文献检索实践课程教学的创新[J]. 实验室研究与探索, 2012, 31(2): 115-118.
Zhang Jia, Dou Lihua, Chen Jie. Teaching innovation of science and technology literature retrieval[J]. Research and Exploration in Laboratory, 2012, 31(2): 115-118.

[6] 邹永利, 何侃, 徐健. 文体特征在网络学术文献检索中的意义与应用[J]. 情报理论与实践, 2008, 31(4): 594-597.
Zou Yongli, He Kan, Xu Jian. The significance and application of stylistic features in network academic literature retrieval[J]. Information Studies: Theory & Application, 2008, 31(4): 594-597.

[7] 张永宏, 胡立耘.文献检索在编辑工作中的应用[J]. 编辑学报, 2001, 13(3):158-160.
Zhang Yonghong, Hu Liyun. Application of knowledge of bibliography to editing[J]. Acta Editologica, 2001, 13(3):158-160.

[8] 黄晓鹏,李树民, 廉立军. 我国高等院校文献检索教学研究文献分析[J]. 现代情报, 2009, 29(3): 222-225.
Huang Xiaoli, Li Shumin, Lian Lijun. Literature analysis of literature retrieval teaching research in Chinese university[J]. Journal of Modern Information, 2009, 29(3): 222-225.