基于语句相似度的网页标题抽取方法*

李国华 咎红英

郑州大学信息工程学院，河南 郑州 450001

Email: liguohuahao@126.com iehyzan@zzu.edu.cn

摘 要：目前网页标题的抽取方法大多利用 HTML 结构和标签特征生成特定规则进行抽取，但是这些方法只考虑到了 HTML 的统计特点，没有考虑标题与正文信息之间的关系。本文提出一种基于相似度的网页标题抽取方法，充分利用了网页标题与正文信息之间的关系，通过计算两两“单位”之间的相似度和对应的权值，并引入 HITS 算法模型对权值进行调整，根据特定的选取方法抽取出真实标题。实验结果表明，该方法不仅对“非标准网页”的抽取达到满意的效果，而且对“标准网页”具有较高的泛化能力。

关键词：网页标题抽取，相似度，HITS 算法，Web 信息抽取

Title Extraction from HTML Documents based on Similarity

Li Guohua, Zan Hongying

School of Information Engineering,Zhengzhou University, Zhengzhou, Henan 450001

Email: liguohuahao@126.com iehyzan@zzu.edu.cn

Abstract: Most of the algorithms for title extraction from HTML documents are based on the structure of HTML document or features of label, then generate the rules in this way. They make use of the characteristic of statistic, but do not consider the correlation between the title and the main body. This paper addressed a method of title extraction from HTML documents based on similarity. The method make full use of the correlation between the title and the main body. We caculate the similarity between units and introduce HITS algorithmic model to adjust the unit's weight, then extract the "real" title in a series of steps. Experimental results show that this method performs well for "nonstandard HTML document" and has good generalization ability for "standard HTML document".

Key words: Title Extraction, Similarity, HITS algorithm, Web Information Retrieval

1 引言

网页文档作为互联网信息的一种载体，人们通过网页文档可以发布和获取各种各样的信息。随着网络信息量的与日俱增，互联网上的海量信息在丰富了人们信息来源的同时，也给人们获取感兴趣的信息带来了困难。面对海量的信息，如何有效地抽取网页文档中的数据，与人们更有效快捷的获取相应信息的要求直接相关。

本文提出了一种基于相似度计算方法的网页“真实”标题抽取方法。我们定义：与文档信息相关的标题为“真实标题”，与正文不相关的标题为“虚假标题”；相应的网页定义为“标准网页”和“非标准网页”；“单位”定义为 HTML 文档抽取出的文本信息的独立句子或段落。

网页标题是一篇网页所要表达信息的最简明扼要的概述，它对于网页信息的处理及应用（比如搜索引擎、聚类和分类）有很大的意义。大多数情况下我们可以通过 HTML 文档中的<title>和</title>标签准确的获得“真实”标题，但有些时候人们却不经意地将“真实标题”表达在自定

* 基金项目：本文研究工作受到国家自然科学基金（项目号 60970083）及国家社会科学基金(09BTQ027)的资助。

义的 HTML 标签中，而在<title>和</title>标签中填写的是“虚假标题”。这会使通过网页上显示的标题进行查找资源的人们被迫错失一些重要的信息来源。

区别于现有的网页标题抽取方法，我们通过对网页进行预处理，将原始网页中的文本信息表示成由多“单位”组成的文档，文档中不包含 HTML 的任何属性标签。然后比较两两之间的相似度，通过一系列计算步骤和方法，最终抽取出“真实标题”。

实验表明我们提出的方法在“标准网页”和“非标准网页”的数据集上都能取得较好的效果，并且可以成功地应用于我们的搜索平台[1]。

本文以下部分的组织结构是：第二部分介绍相关研究；第三部分详细介绍基于该方法的相关内容；第四部分给出了本方法实验结果和说明；第五部分给出总结及下一步的工作。

2 相关研究

Web 信息抽取方法目前大多是基于规则，一是利用自然语言处理技术的词法、子句结构、短语和子句间的关系建立基于语法和语义的抽取规则。典型的系统有 SRV[2]、WHISK[3]等。该方法是将 Web 文档视为文本进行处理，较适合含有大量非结构化文本的 web 页面。二是利用机器学习方式生成基于定界符的抽取规则[4-8]。规则的获取需要训练手工标注的样本实例。典型的系统有 Stalker、SoftMealy[9]和 WIEN[10]等。与基于自然语言理解方式相比较，该方法仅仅使用语义项的上下文来定位信息，没有使用语言的语法约束。这两种基于规则的抽取方法都需要训练样本，自动化程度低[11]。

一般来说，同一个单位内的网页结构基本相似，或干脆使用同一套网页模板，文献[5]和文献[12]也考虑了同样的策略。基于此方法，我们的校内搜索引擎[1]前期版本是通过手工制定规则来获取网页的标题，但是这种方法即使是在小范围内也需要耗费很大的人力。

文献[12]和文献[13-15]通过将网页分析为 DOM 树，然后从 DOM 树中提取出信息，文献[13]将网页分析为 DOM 树，然后从中提取出含有特征属性的单位，结合自定义的各种 HTML 特征的重要程度来提取标题。文献[16]也利用 HTML 的主要特征研究对 web 信息检索的作用。文献[17]学习一种发现网页中重要的块的模型。文献[18]通过基于网页布局的相似度进行 web 论坛数据的抽取。

虽然很早便有相类似的工作应用在自动文摘的研究中[19-22]，本文的研究工作和文献[22]相类似。据我们所知，目前为止，还没有利用句子之间的相似度为基础来进行网页标题抽取的研究。

3 网页标题的抽取

3.1 网页文档预处理

计算句子之间的相似度，首先需要将 HTML 文档中含有的信息转换为文本文档表示。此处，我们使用 Nekohtml[23]开源工具包进行转换。Nekohtml 是一个 Java 语言的 HTML 扫描器和标签补全器，借助 Nekohtml 我们可以解析 HTML 文档并得到 HTML 文档包含的所有纯文本信息。

在转换过程中，对于 Element 节点，我们增加“\n”为获得该节点信息的结束标志，从而在转换完成后，我们可以对整个纯文本信息以“\n”进行划分，将经过划分后的段落或句子等同定义为一个“单位”。正文中的噪声比如广告、导航信息或相关链接会分别以一个“单位”对待；

网页中的真实标题也会独立成为一个“单位”；正文信息则由一个或多个“单位”组成。

3.2 相似度的计算

考虑到标题信息为网页正文信息的高度概括，其长度与正文信息的长度相比差距较大，所以选择利用正向迭代最细粒度切分算法分词后的公共子词语方式计算单位间的相似度。“正向迭代最细粒度切分算法”分词方法：比如“郑州大学”分词后为：“郑州大学”、“郑州”、“大学”。

计算两个单位 unit_1 和 unit_2 间的相似度方法如下：

$$\mathrm{Sim}(\mathrm{unit\_1}, \mathrm{unit\_2}) = \mathrm{Sim}(\mathrm{set\_1}, \mathrm{set\_2}) = \frac{sameCT * sameCT}{\log(size(set\_1) + size(set\_2))} \qquad (1)$$

其中 set_1 和 set_2 分别为需要计算的两个单位 unit_1 和 unit_2 经过迭代分词后的词语集合。如果集合中出现相同词语，只保留一个词语，且词语的数值为集合中词语出现的次数，set 内的数据结构表示为<word, count>，word 为词语，count 为 word 出现的次数。

sameCT 为 set_1 和 set_2 两个集合的共同词语数之和，和的值等于共同词语的次数相加。size(set)表示 set 集合的长度，sameCT 的计算公式如下：

$$sameCT = \sum CT1(Word_i) + \sum CT2(Word_i) \quad Word_i \in \mathrm{set\_1} \text{或} Word_i \in \mathrm{set\_2} \qquad (2)$$

3.3 权值的计算

根据公式（1）计算出的两两单位之间的相似度，可以得到一个单位的权值计算公式：

$$\mathrm{Weight}(\mathrm{unit\_i}) = \sum_{j=0}^{N} Sim(unit\_i, unit\_j) \qquad (i \neq j) \qquad (3)$$

unit_i 为需要计算权值的单位。Sim(unit_i, unit_j)为 unit_i 与 unit_j 的相似度。N 为文档中的单位的总数目。

3.4 权值的调整

HITS 算法通过两个评价权值——内容权威度（Authority）和链接权威度（Hub）来对网页质量进行评估。

本文将其思想应用到文本文档中的各个单位之间，首先将文本文档表示成图 G。图 G 的各个顶点分别对应各个单位；顶点之间的边是否存在取决于顶点对应的单位之间相似度的大小，如果相似度的值等于 0，则顶点之间不存在边；边的权值大小为相似度的值，值大于 0；顶点的初始权重为公式（3）计算出的权值大小。

根据图 G 的定义，我们对公式（3）计算出的单位的权值进行调整：

$$Weight'(\mathrm{unit\_i}) = \mathrm{Weight}(\mathrm{unit\_i}) * \mathrm{linkCT}(\mathrm{unit\_i}) \qquad (4)$$

其中 Weight(unit_i)为 unit_i 的初始权重，即公式（3）计算出的权值。linkCT 为图 G 中单位 unit_i(unit_i)对应顶点的度。

公式(4)表明，一个顶点的度越大，其对应的单位的重要性也就越大。

3.5 标题的选取步骤

我们将整篇文本文档以“\n”划分成多个单位，并通过计算后，表示成 Collection<<u nit_i, weight_i>>sortList。以下是标题选取的步骤：

1，首先对 sortList 按照文档中的单位 unit 的权值$Weight'$(unit)进行升序排序；

2，计算所有顶点的度数和 TTCT 以及权值大于等于∂的顶点总个数 PCT：

$$TTCT = \sum linkCT(unit\_i) \quad Weight'(\text{unit\_i}) >= \partial \tag{5}$$

其中∂为可定义的参数值，实验测试取∂值为 0.1 比较合适。

$$PCT = \sum \begin{cases} 1 & Weight'(\text{unit\_i}) >= \partial \\ 0 & Weight'(unit\_i) < \partial \end{cases} \tag{6}$$

3，计算平均度的阈值 aveCT：

$$aveCT = \frac{TTCT}{2 * PCT} \tag{7}$$

其中 aveCT 为用于控制权值过小的单位。判断条件为：如果 linkCT(unit_i) < aveCT，则不考虑将该单位作为候选的标题。

4，经过步骤 1，2，3 的计算：

第一：我们选取 sortList 中序号 idx 较小的两个单位作为候选标题。单位的序号 idx 定义为该单位在文本文档被划分为多个单位中相对应的索引序号。这里选取原则为“真实标题”往往出现在网页的顶部区域，其索引序号较小。

第二：比较两个候选单位的权值，选取权值较大的单位作为我们抽取出的“真实标题”。

4 实验

4.1 数据集的选取

为了验证提出的方法的有效性，我们从校内搜索引擎[1]抓取的网页中选取。通过人工制定规则获取真实标题，并校对验证真实标题的正确性，剔除出现乱码和全英文的网页后，一共 23709 篇“非标准网页”用于“非标准网页”标题抽取的实验。

由于“标准网页”的“真实标题”能够很准确的出现在 HTML 的<title>和</title>标签域里面，我们只需要通过分析 title 域便可以得到“真实标题”。

同时，为了验证提出的方法的泛化能力，本文从 web 上的 7 个站点（北方网、新浪网、搜狐网、中华网、新民网、网易网、艾瑞网）的子栏目利用爬虫抓取了 3000 篇“标准网页”，并且从郑州大学内部网中抓取了 250 篇“标准网页”，一共 3250 篇“标准网页”用于“标准网页”标题抽取实验。

“标准网页”数据集的来源及选取的网页篇数，请见表 1。

表 1 “标准网页”数据集来源及篇数

	种子网址	篇数
郑大内部网	http://www2.zzu.edu.cn/history/	250
北方网	http://news.enorth.com.cn/gn/wjwd/	300
新浪网	http://news.sina.com.cn/china/	300
搜狐网	http://news.sohu.com/guoneixinwen.shtml	300
中华网	http://news.china.com/zh_cn/domestic/index.html	300
新民网	http://news.xinmin.cn/rollnews/	300
网易网	http://news.163.com/special/00013C0O/guojibjtj.html	500
艾瑞网	http://news.iresearch.cn/list468/	1000

4.2 标题抽取的评测方法

我们使用准确率作为标题抽取结果的评估。准确率的计算公式为：

$$\text{准确率}=\frac{\text{标题抽取正确的HTML文档数目}}{\text{总的HTML文档数目}}*100\% \tag{8}$$

同时，利用本方法抽取出的标题和“真实标题”的近似程度超过阈值 β 时，我们判定为抽取正确。此处近似值的计算方式为：

$$\frac{sameCT}{size(title\_extracted)}>=\beta \tag{9}$$

sameCT 为抽取出来的标题 title_extracted 和“真实标题”的共同子词语数；size(title_extracted)为抽取出来的标题的长度；β 等于 0.6。

参数 β 的选取主要因为网页的标题中，发布人通常会在网页的标题后面加上信息来源，比如：“美国冒险家徒手登上海波 3900 米高峰(组图)-冒险-北方网-科技无限”。

4.3 “标准网页”标题抽取实验

表 2 “标准网页”标题抽取的结果

	网页篇数	得到的正确篇数	准确率
郑大内部网	250	250	100.00%
北方网	300	300	100.00%
新浪网	300	290	96.67%
搜狐网	300	297	99.00%
中华网	300	300	100.00%
新民网	300	300	100.00%
网易网	500	498	99.60%
艾瑞网	1000	1000	100.00%
total	3250	3235	99.54%

从表 2 中我们可以看出，该方法对于 web 网上的网页抽取准确率很高，泛化能力可以得到保证。经过对由方法抽取的标题与正确标题进行对比并观察网页发现，抽取错误的网页特征主要集中表现为：类型一、网页是链接导航型的网页，即网站的子分类栏目或某个专题的索引页面，网页中正文信息过于分散；类型二、网页新闻的标题为使用近似语义概括的标题，由于本方法没有进行同义词扩展，所以对于这类网页，抽取出的效果也不是很好。

4.4 “非标准网页”标题抽取实验

表 3 “非标准网页”标题抽取结果

	网页篇数	得到的正确篇数	准确率
郑大内部网	23709	21485	90.62%

从表 3 中我们可以看出，该方法对于较大数据集的“非标准网页”处理性能仍然较好。经过对由方法抽取的标题与正确标题进行对比并观察网页发现，抽取错误的网页特征除有“标准网页”中出现的两种类型的错误外，还表现为：类型三、网页正文信息表达了多个主题，对于这种网页，本方法抽取出的结果大都是其中的一个子主题的标题；类型四、网页为图片或内容为表格或文件下载，文字信息很少。

四种错误类型的网页的统计的数据见表 4。

表 4 标题抽取错误的结果中——属于四种错误类型的网页所占篇数

	类型一	类型二	类型三	类型四	总共
非标准网页（篇数）	190	——	104	60	354

注：类型二“——”表示无法准确判断网页是否归属类型二。

以上对“标准网页”和“非标准网页”的标题抽取实验数据显示，本文提出的方法对于抽取“非标准网页”的“真实标题”性能良好，同时对互联网网页的泛化能力较高。

5 结论与展望

本文提出了一种基于相似度的网页标题抽取方法，区别于利用 HTML 结构和标签特征的标题抽取方法，并取得了令人满意的抽取效果。实验表明本文提出的方法不仅可以满意地实现对“非标准网页”的抽取，而且对“标准网页”有较好的泛化能力。下一步将考虑改进相似度比较方法以及更深入的挖掘 HITS 模型对权值的调整等工作。

参考文献

[1] 郑州大学校内搜索引擎. http://www.jiubukan.com.

[2] Freitag D. Machine Learning for Information Extraction in Informal Domains. Machine Learning, 2000.

[3] Soderland S. Learning Information Extraction Rules for Semi-structed and Free Text. Machine Learning, 1999.

[4] Yipu Wu, Xuejie Zhang, Qing Li, Jing Chen. Title Extraction from Loosely Structured Data Records. In Proceedings of the Seventh International Conference on Machine Learning and Cybernetics, 2008.

[5] Crescenzi, V., Mecca, G. and Merialdo, P. Roadrunner: Towards Automatic Data Extraction from Large Web Sites. In

Proceedings of the Twenty-seventh International Conference on Very Large Databases(VLDB2001), 2002.

[6] Chidlovskii, B., Ragetli, J., and de Rijke, M. Wrapper Generation via Grammar Induction. In Proceedings of the Eleventh European Conference on Machine Learning(ECML2000), 2000.

[7] Crescenzi, V., Mecca, G. and Merialdo, P. Wrapping-Oriented Classification of Web pages. In Proceedings of the 2002 ACM Symposium on Applied Computing(SAC-2002), 2002.1108~1112.

[8] Craven, T.C. HTML Tags as Extraction Cues for Web Page Description Construction, Informing Science Journal, Volume 6, 2003.

[9] Hsu C N, Dung M T . Generating Finite-State Transducers for Semi-Structured Data Extraction from the Web. Information Systems, 1998.

[10] Kushmerick N, Weld D S. Doorenbos R. Wrapper Induction for Information Extraction. 15th International Joint Conference on Artificial Intelligence (IJCAI-97), Nagoya, 1997.

[11] 李猛. 基于DOM的Web信息抽取技术的研究与实现[D], 2008:5~6.

[12] Kosala, R., Bruynooghe, M., Bussche, J.V. and Blockeel, H. Information Extraction from Web Documents Based on Local Unranked Tree Automaton Inference, InProceedings of the Eighteenth International Joint Conference on Artificial Intelligence(IJCAI-2003), 2003.

[13] Yunhua Hu, Guomao Xin, Ruihua Song, Guoping Hu, Shuming Shi, Yunbo Cao, and Hang li, "Title Extraction from Bodies of HTML Documents and its application to Web Page Retrieval", Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval,2005.250~257,.

[14] Breuel, T.M Information Extraction from HTML Documents by Structural Matching, In Proceedings of the Second International Workshop on Web Document Analysis(WDA2003), 2003.

[15] Reis, D., Golgher, P., Silva, A. and Laender, A. Automatic Web News Extraction Using Tree Edit Distance. In Proceedings of International WWW Conference(WWW-2004),2004.

[16] Zhang, M., Song, R. and Ma, S. DF or IDF?On the use of HTML primary feature fields for Web IR. In Proceedings of the Teelfth International World Web Conference(WWW2003), 2003.

[17] Song, R., Liu, H., Wen, J.-R. and Ma, W.Y. Learning Block Importance Models for Web Pages, In Proceedings of International WWW Conference(WWW-2004), 2004.

[18] 王允, 李弼程, 林琛. 基于网页布局相似度的Web论坛数据抽取.《中文信息学报》. Vol. 24 No.2, pp 68-75, 2010年2月.

[19] GSalton, A. Singhai, M. Mitra, C.Buckly. Automatic text structuring and summarization [A]. In advances in Automatic Text Summarization [C], Eds. I. Mani and M.T.Maybury. The MIT Press,1999:62~70.

[20] Jae-Hoon Kim, JoonHong Kim , Dosam Hwang, 2000. Korean Text Summarization Using an Aggregate Similarity [A]. The 5th International Workshop on Information Retrieval with Asian Languages [C]. Hong Kong , September 30 to October 3, 2000.

[21] 张奇, 黄萱菁, 吴立德. 一种新的句子相似度度量及其在文本自动摘要中的应用.《中文信息学报》. Vol.19 No.2, pp 93-98,2005年2月.

[22] Rada Mihalcea. Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization. In Proceedings of the Conference and Workshops of ACL-2004. Barcelona.

[23] Nekohtml. http://nekohtml.sourceforge.net/.