基于关键词聚类和节点距离的网页信息抽取*)

邓健爽　郑启伦　彭　宏　林旭东

(华南理工大学计算机科学与工程学院人工智能实验室　广州 510641)

摘　要　大部分网页信息抽取方法都针对特定的网站,例如基于网站抽取规则和基于训练网页样例的方法。这些方法在某一个网站上可以很好地应用。但当遇到新的网站时,必须人为地增加抽取规则或者提供新的训练网页集。而且,当网站的模版改变时,也要重新设计这些规则或重新输入训练网页集。这些方法难以维护,因此不能应用到从大量不同的网站上进行信息抽取。本文提出了一种新的网页信息抽取方法,该方法基于特定主题的关键词组和节点距离,能够不加区分地对不同的网站页面信息自动抽取。对大量网站的网页进行信息抽取的实验显示,该方法能够不依赖网页的来源而正确和自动地抽取相关信息,并且已经成功应用到电子商务智能搜索和挖掘系统中。

关键词　聚类,信息抽取,机器学习,节点距离

Web Pages Information Retrieval Based on Keywords Cluster and Node Instance

DENG Jian-Shuang　ZHENG Qi-Lun　PENG Hong　LIN Xu-Dong

(Department of Computer Science, The South China University of Technology, Guangzhou 510641)

Abstract　Many Web information retrieval methods are related to special Web sites, for example, the method based on extracting rules and the one based on training page samples. These methods can do well in a Web site but fail in the others without adding new rules or inputting new training pages manually. Furthermore, if the template of the Web site is changed, it has to redesign the extracting rules or re-inputting the training pages. It is hard to be maintained and used to extract information from large number of different Web sites. In the paper, there is a new method which can extract the useful information from the different sites automatically based on the keywords of a certain topic and the distance of the nodes. Experimental evaluation on a large of Web pages from different Web sites indicates that this method correctly and automatically extracts the information ignoring which Web sites the pages come from. This method has been applied to the system of intelligent searching and mining of electronic business successfully.

Keywords　Cluster, Information retrieval, Machine learning, Instance of node

1　引言

互联网上包含着海量信息,但是一般用户难以找到自己真正所需的信息。为了解决用户在互联网上盲目查找信息的问题,发展了搜索引擎技术和信息抽取技术。其中一般的搜索引擎可以找到与用户信息相关的文档,但具体的信息还要用户亲自打开文档,进行判断和提取。而信息抽取技术可以为用户自动地从文档中提取有用的信息,并且可以把互联网上分散的、杂乱无章的同类信息组织起来为用户提供服务。互联网上大部分的信息都是存储在半结构化的网页里,如何从网页里提取有用信息成了一项重要的研究课题。基于模板的信息抽取[1~3]和基于机器学习的信息抽取[4~6]是两种典型的网页信息抽取方法。基于模板的信息抽取技术能够很好地把有用的信息抽取出来,准确率高,但是每个模板只能应用到同一个网站或者网页中,对于大量网站的信息抽取,要产生大量模板,而且互联网的信息更新很快,要求相应频繁地更新模板,同时要求模板的编写人员对网页有深入的了解和相关知识,这就大大增加了维护的代价。基于机器学习的信息抽取要求用户输入一定数量的训练网页,然后通过归纳、分类、聚类等人工智能和数据挖掘的方法进行信息抽取,它减少了用户维护模板的代价,但是也要用户准备大量样本网页,而且抽取的准确率一定程度上与样本网页的选取和数量有关。同时,不同网站的信息组织结构不同,也需要输入不同的样本网页,大大增加了用户的负担和减少在不同网站之间的适用性。有没有网页信息抽取方法能够不区分网页来自什么网站而能够自动正确地抽取相关重要内容?有没有网页信息抽取方法与网页的模版无关,当网页更新时也能自动正确地抽取信息?为了解决这些问题,我们提出了一个新的基于主题关键词组和节点距离的网页信息抽取方法,能够自动从大量网站上进行信息抽取。

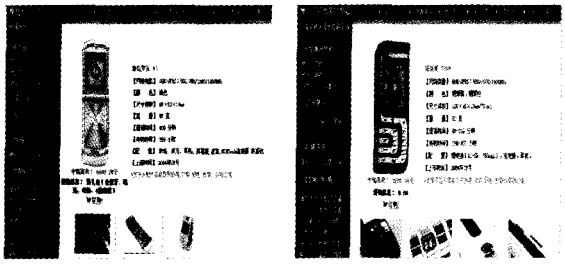
2　基于关键词和节点距离的信息抽取

与搜索引擎没有特定的主题不同,例如 GOOGLE[7],信息抽取一般是面向某个特定的主题。经过分析,我们发现,互联网上同一网站一般采取相同的模板生成同类主题,而不同网站相同主题的组织形式也有很多相似的地方,如图 1。

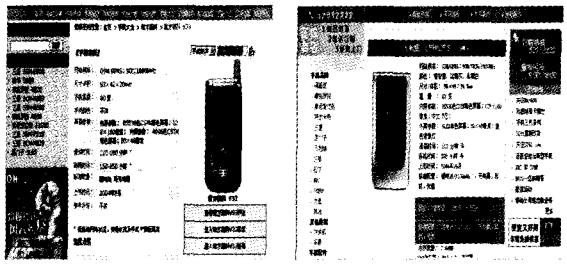
我们以电子商务网站中关于手机产品信息的抽取为例进行说明。从图 1 中可以发现,尽管不同网站中关于手机的信

*)广东省科技攻关项目(2005B10101033)(A10202001)、广州市科技攻关项目(2004Z2-D0091)。**邓健爽**　博士研究生,主要研究方向为人工智能、网络智能搜索和数据挖掘;**郑启伦**　教授,博士生导师,主要研究方向为人工智能与海量数据处理相关的智能计算技术及其应用;**彭　宏**　教授,博士生导师,主要研究方向为数据挖掘;**林旭东**　博士研究生,主要研究方向为网络智能搜索和数据挖掘。

息组织结构不同，但是它们都包含类似的手机介绍信息，例如颜色、尺寸、产地、质量、价格等和手机的图片，并且集中一起放在最中央的位置。



(a) 相同网站上不同的网页



(b) 不同网站上的网页

图 1　网页样式的比较

通过上述分析，本文提出了一种全新的不区分网站的网页信息抽取方法——基于关键词和节点距离的网页信息抽取。该方法不需要生成抽取模板，而是通过分析相同主题网页的特征，产生关键词组，再利用这些关键词在网页上的分布，通过聚类方法[8]确立关键信息块[9,10]，然后对关键信息块进行信息抽取，并且利用到关键信息块的节点距离抽取其他重要信息。图 1 中关键信息块就是网页中手机的介绍信息块。在我们的电子商务智能搜索及挖掘系统中应用该方法成功地对 30 多个电子商务网站大约 10000 个手机网页进行手机型号、手机价格和手机图片的抽取。实验证明，该方法还具有良好的扩展性，能够方便地移植到对其他商品信息的抽取。用户所要做的就是根据相关的主题或商品，选择关键词组。不同网站间的信息抽取对用户来说是完全透明的，网站的更新和改动也不会影响抽取结果。

3　基本步骤分析

基于关键词聚类和节点距离网页信息抽取方法的基本步骤(流程如图 2)是：

1)确定抽取主题相关的关键词组。

2)标准化源文件并且根据网页源文件建立结构树。

3)利用聚类方法查找包含大部分关键词的子树的根节点，该根节点就是我们所需要的关键信息块节点(简称关键节点或中心节点)。

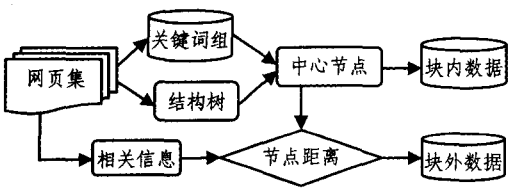


图 2　系统流程图

4)提取关键信息块里的信息。

5)对于关键信息块外的信息，计算该信息所在节点与关键节点之间的节点距离。

6)通过权值计算确定信息，并最终抽取该信息。

3.1　确立关键词组

在同一领域里不同网站介绍同一商品的网页，都包含一些与商品有关的相同或相似的关键词。而这些关键词通常放在重要的信息块里面，可以通过确立和定位这些关键词来确定和抽取关键信息块。关键词组与一定的领域知识和商品密切相关，我们需要为不同的商品建立不同的关键词组。对于某一商品的关键词组的选取，要使得该关键词组覆盖尽可能多的网站，同时每个网站包含其中一定数量的关键词。但是并不表示关键词越多越好，太多的关键词会造成干扰。实际上，一般关键词组包含关键词的数量都比较少，而且一旦建立了某商品对应的关键词组，即使网页的样式更新和应用到新的网站中去也不需要再修改。我们可以建立一个商品关键词库，包含多种商品的关键词组。当要在电子商务网站抽取某种商品信息时，直接去商品关键词库取得相应商品的关键词组。该方法具有很强的适应性和扩展性。

对于某类商品，获取一定数量的相关网页，对这些网页进行文本分析，提取频繁关键词作为商品的相关关键词。由于属于同一网站的网页往往是经过相同的模版生成的，而不同网站都有各自的模版，因此针对网页是否属于同一网站进行相应不同的处理。本文采用分层抽取的思路来挖掘商品相关关键词，如图 3。

算法：

输入：某商品一定数量的样本网页；网站内频繁 λ_1，网站间频繁度 λ_2
输出：该商品的相关关键词集
步骤：
对样本网页根据所在的网站进行分类；
　　同一网站的网页集频繁关键词提取{
　　抽取网页纯文本；
　　调用分词器对网页文本进行分词处理；
　　计算网站内关键词的词频 f_1；/ * f_1 = 网站内包含该关键词的样本网页数 / 该网站的总样本网页数 * /
　　返回网站内的词频大于 λ_1 的关键词集；
}
网站间频繁关键词提取{
计算网站间关键词的词频 f_2；/ * f_2 = 包含该关键词的样本网站数/所有样本网站数 * /
　　返回网站间的频繁度大于 λ_2 关键词集；
}

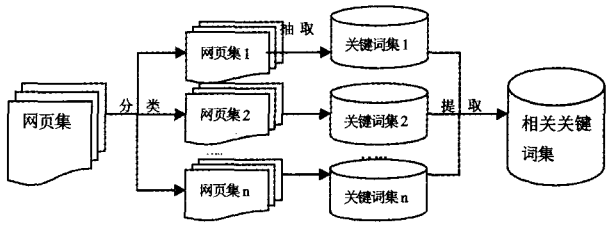


图 3　关键词组的提取

这里以在电子商务网站中抽取手机功能信息、图片和价格为例进行说明。经过对 20 个电子商务网站中关于手机的 200 个网页进行分析，其中 $\lambda_1 = 90\%$，$\lambda_2 = 30\%$，得出表 1 的结果。

表 1 中 $\overline{f_1}$ 表示关键词站内词频大于 λ_1 的所有网站的 f_1 的平均值。以上方法所建立的关键词组都是一些很有代表性的关键词，但是还有很多相类似的或同义的关键词，我们不在关键词组里表示，而增加一个同义词库，在程序运行过程中，进行关键词的同义词扩展。例如对“价格”关键词进行查找时，我们同时对“市场价”、“价钱”、“参考价”、“¥”、“售价”这些同义词进行查找，把结果统一按照“价格”关键词来处理。

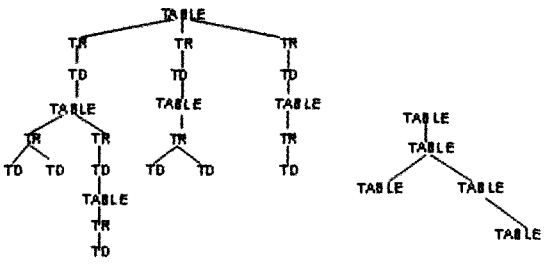
这里我们所选的这 12 个关键词能够覆盖全部 20 多个电子商务网站，而且每个网站都至少包含 4 个其中的关键词。我们就通过查找这 12 个关键词在网页里的位置，从而确定关键信息块。

表 1　手机关键词组的选取

关键词	尺寸/体积	重量	颜色	通话时间
$\overline{f_1}$	100%	100%	99%	100%
f_2	100%	100%	75%	85%
关键词	网络	待机时间	配置	显示屏
$\overline{f_1}$	100%	99%	100%	98%
f_2	80%	80%	68%	55%
关键词	型号	上市时间	价格	规格
$\overline{f_1}$	100%	100%	100%	98%
f_2	32%	45%	64%	55%

3.2　构建网页结构树

由于 HTML 文件的层次和嵌套等特点，可以通过〈tag〉标记建立网页的结构树，对网页的内容和结构进行分析。传统的网页结构树方法，大部分的标记（如〈table〉〈tr〉〈td〉〈p〉等）都在结构树作为一个树节点，从而组成一棵复杂的多叉树。虽然某些标记对内容和结构的分析具有一定作用，但是也有很多标记实际上毫无意义，成为干扰信息。而且建立一棵这样的树在时间和空间上的花销都比较大。为了简化结构树，我们只是对源文件中〈table〉标记进行分析并且构造一棵以〈table〉标记为节点的二叉树。树的所有节点都是〈table〉节点，每个节点包含重要信息，从而使树的结构简化了，但信息却没减少。树中左节点是子节点，右节点是兄弟节点，如图 4。



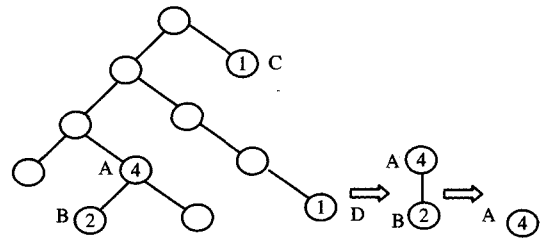
(a)传统结构树　　(b) 我们简化的结构树

图 4　HTML 源文件的页面结构树

3.3　确定和抽取关键信息块

通过关键词组的分布来定位关键信息块。首先查找词组中各个关键词在源文件中的位置，并且在页面结构树中确定包含该关键词的节点，从而得到包含关键词的树节点集。通过聚类的方法找到最大子节点集，然后求得包含该子节点集的最小子树，该子树的根节点就是包含关键信息块的关键节点。由于关键词也有可能出现在关键信息块以外的其他地方，造成干扰，因此我们要通过聚类方法找到最大子节点集，去除这些干扰。找到关键节点，我们就可以根据节点的信息，找到关键信息块起始标记〈table〉〈/table〉的位置，提取关键信息块内容。

如图 5 所示，节点里面的数字表示该节点所包含的关键词数目，节点旁边的字母标识该节点。一共有 4 个节点包含关键词，其中 A 节点包含 4 个、B 节点包含 2 个、C，D 节点各包含 1 个。我们对这 4 个节点通过聚类算法分类：AB 为一类，C 为一类，D 为一类。通过分析，我们认为 C，D 为干扰节点，只提取 A，B 节点。然后查找包含 AB 节点的最小子树。由于节点 A 的右节点实际是 A 的兄弟节点，我们把它从以节点 A 为根节点的子树中删除，得到最小子树，如图 5(b)，再提取最小子树的根节点 A，如图 5(c)。最后根据节点 A 对应〈TABLE〉〈/TABLE〉的位置提取包含的内容，抽取关键信息。



(a) 分析树　　(b) 最小子树　　(c) 关键节点

图 5　关键节点的确立

3.4　提取关键信息块以外的信息

在电子商务网站里，像图片、价格等其他一些重要信息都可能不是包含在关键信息块里的，但是我们可以通过关键节点和节点距离找到这些信息。

本文定义了一个新的概念——节点距离。节点距离反映两个节点在结构树里的紧密程度：

$$l=\sqrt{(d_1-d_2)^2+(w_1-w_2)^2} \tag{1}$$

其中 d 是节点在结构树中的深度，这里深度的定义有所不同。由于我们构建的二叉结构树中右节点实际上是兄弟节点，因此所有节点和它的右节点的深度是相同的，而左节点的深度加 1。w 是节点在结构树中的宽度，和深度相反，右节点宽度加 1，左节点宽度不变。

关键信息块一般包含商品的介绍和说明信息，它可以作为我们抽取其他信息的依据。我们以关键节点为中心，通过其他节点与之的节点距离来计算该节点包含信息的重要性。例如我们要抽取手机图片，首先查找源文件里包含的所有图片文件，并且计算它们所在节点与关键节点的节点距离，通过该节点距离与其他因素一起计算图片的权重。实验证明，通过节点距离可以很好地筛选出所需要的重要信息，如图片、价格等内容。为了网站的美观，网页上一般包含大量的图片。正是由于源文件上图片信息的异常丰富，因此要正确抽取关于手机的图片，单单靠节点距离还是不够的，我们添加了更多的限制因素。关于图片的大小、形状、位置等因素都考虑在内。我们用到下面公式：

$$W=\frac{k_1}{1+l}+\frac{k_2 S}{S}+\frac{k_3(h-w)}{w} \tag{2}$$

式中 W 是图片的权值，k_1，k_2，k_3 是权值调整系数，l 是节点距离，s 是图片的面积大小，S 是源文件里面积最大的图片的大小，h 是图片的高度，w 是图片的宽度。因为经过观察，绝大部分的网页中手机的图片都是竖立的长方形，所以我们增加上式的第 3 项来把这种图片形状特性反映到我们手机图片的权值计算公式里去。为了去掉大部分小图片的干扰，在这里我们还为 h 和 w 设置了一个阈值 λ，当 h 或 w 小于这个阈值 λ 时，我们设图片的权值为 0。对于非图片的其他信息，如价格、型号，我们在权值计算中相应增加一些其他因素。

4 实验结果

我们通过对30多个电子商务网站中关于手机的网页进行手机型号、价格、图片等信息的抽取[11]。其中的 k_1，k_2，k_3，λ 分别取0.5，0.3，0.2，50。这个实验是在CPU为Intel CeleronⅢ 1.3GHz，256内存，Windows Sever 2003操作系统的机器上进行。

实验结果如表2。

表2 手机网页信息抽取

Part1

网站	测试网页数量	图片正确率	价格正确率	型号正确率	时间(ms)
6226.com.cn	170	100%	100%	100%	174
www.sh123.net	166	100%	100%	100%	474
www.18900.com	19	100%	100%	100%	967
www.yephone.com	497	100%	100%	99%	375
kinfic.com.cn	29	100%	100%	100%	123
www.one2free.com.cn	30	100%	94%	96%	469
www.5291.com	77	99%	100%	99%	98
www.younet.com	980	100%	100%	98%	231
www.my8848.com	87	98%	96%	98%	1271
www.139shop.com	269	99%	100%	97%	97
www.e8888.net	102	98%	97%	99%	250
www.1380138.com	450	100%	100%	100%	200
www.shooji.net	995	100%	100%	100%	374
Mobile.online.sh.cn	131	100%	100%	99%	135
www.519shop.com	324	100%	100%	98%	152
www.mt365.net	79	100%	100%	100%	732
Mobile.cnool.net	35	97%	100%	95%	817
bj816.q88.net	82	98%	98%	98%	234
www.sh169.com	13	100%	100%	100%	321
www.51buy.cn	99	100%	99%	99%	187

Part2

网站	测试网页数量	图片正确率	价格正确率	型号正确率	时间(ms)
vlongbiz.com	21	95%	97%	100%	263
bpcall.com	89	100%	100%	100%	344
www.21cn.com	225	99%	100%	100%	372
www.558shop.com	691	100%	100%	100%	176
shopimg.6688.com	56	100%	100%	100%	93
sas686.com	18	100%	100%	100%	141
www.tom.com	450	99%	99%	99%	109
www.dongdong.com.cn	215	99%	99%	99%	131
www.dd5600.com	204	100%	100%	100%	513
www.imobile.com.cn	262	99%	98%	97%	94
Mobile8848.com	65	97%	100%	100%	464
www.sjinfo.net	87	100%	100%	100%	232
www.163.com	189	99%	99%	99%	347
83222.com	360	100%	100%	100%	157

·适应性

表中PART1部分是我们分析手机关键词组所用到的网站，PART2部分是我们新增的用来测试的网站。实验证明，该方法遇到新的网站，不需要任何修改，就能自动地从该网站上抽取我们所要的信息，具有很好的适应性。

·扩展性

我们把该方法应用到电子商务网站其他商品上，发觉同样能够成功地抽取该商品的相关信息和图片，有良好的扩展性。

·效率

我们通过从开始建立结构树到成功抽取信息所需要的时间来考察该方法的性能。从结果发现，该方法具有很好的效率。主要是因为我们只是简单地以〈TABLE〉标记来建立分析树，而树的深度一般不超过5，树的建立、查找等操作所需要的时间大大减少了，同时整个过程不需要任何训练学习，进而提高信息抽取效率。

我们已经把实验的结果应用在我们商品智能搜索与挖掘系统的网站 http://202.38.215.1:8080/esearch 上。结果显示，我们的信息抽取方法不但能应用到固定的网站，而且能够应用到大量的不确定的新网站上去。该方法能够提高搜索引擎的智能和推动互联网信息抽取技术的发展。

结束语 本文提出的基于关键词聚类和节点距离的网页信息抽取方法，不需要为不同网站提供不同的抽取模版和训练网页；对于新增的电子商务网站，不需任何修改就能进行信息抽取，具有很好的适应性；对于不同商品的信息抽取，只需要改变相应的关键词组和一些必要的限制因素，具有很好的扩展性，同时也不受网站更新样式的影响，用户维护的工作量少，尤其适合同时对大量的网站进行信息抽取。我们把该方法应用到电子商务智能搜索及挖掘系统中，并且取得良好的效果。

参考文献

1 Crescenzi V, Mecca G. Grammars have exceptions. Information Systems, 1998, 23(8)

2 Hammer J, Garcia-Molina H, Cho J, et al. Extracting semistructured information from the Web. In: Proc. of the Workshop on the Management of Semistructured Data, 1997

3 Huck G, Frankhauser P, Aberer K, et al. Jedi: Extracting and synthesizing information from the web. In: CoopIS, 1998

4 Lerman K, Minton S N, Knoblock C A. Wrapper Maintenance: A Machine Learning Approach. Journal of Artificial Intelligence Research, 2003, 18: 149～181

5 Arasu A, Garcia-Molina H. Extracting Structured Data from Web Pages. In: SIGMOD 2003, San Diego, CA, June 2003

6 Soderland S. Learning information extraction rules for semistructured and free text. Machine Learning, 1999, 34: 1～3

7 Brin S, Page L. The Anatomy of a Large-Scale Hypertextual Web Search Engine

8 Han Jiawei, Kamber M. Data Mining: Concepts and Techniques. China: China Machine Press, 2001

9 Cai Deng, Yu Shi-Peng, Wen Ji-Rong, et al. Block-based Web Search. SIGIR'04, Sheffield, South Yorkshire, UK, July 2004

10 Cai Deng, He Xiao-Fei, Wen Ji-Rong, et al. Block-level Link Analysis. SIGIR'04, Sheffield, South Yorkshire, UK, July 2004

11 Deng Jian-Shuang, Lun Qi, Peng Hong. Information retrieval from large number of Web sites. ICMLC2005, 2005, 4: 2172～2177