计算机科学2002Vol. 29№. 6

Web 文档清洗技术*)

Research on Web Document Cleaning

张 波 王继成 王 强 张福炎

(南京大学软件新技术国家重点实验室 计算机科学与技术系 南京210093)

Abstract Information on Web is expanding rapidly, but the quality differs greatly, which makes Web information retrieval and mining more difficulty. Not only research on the technology of information retrieval and Web mining itself needs to be made, but also cleaning Web documents must be done before Web information retrieval and Web mining. However, the latter is often delegted in most current reseach work. This paper puts forward the concept of Web document cleaning, introduces the role that Web document cleaning plays in Web information processing and the process of Web document cleaning. A rule-based system of Web document cleaning is implemented.

Keywords Web document cleaning, Machine learning, Information extraction

1 引言

随着 Internet 在全球的迅速发展，WWW(World Wide Web)已经发展成为一个包含多种信息资源、站点遍布全球的巨大信息服务网络，为用户提供了一个极具价值的信息源，并已成为世界范围内信息共享和信息传播的最主要渠道之一。WWW 系统一经出现，就得到了迅猛的发展，无论是 WWW 站点数还是 WWW 用户数，都是以每年5～10倍的速度呈指数形式增长。目前仅中国的 Internet 用户就已经达到了2500万。但是随着网络上信息资源的迅速膨胀，WWW 的开放性、异构性和动态性又导致了网络信息检索问题的出现，这主要表现在如下几方面[1]：

(1)WWW 的异构性，决定了其上的信息资源是多种多样的，没有统一的或者结构化的组织形式和存储结构，也没有统一的访问界面，使得用户无法象检索关系型数据库那样对 WWW 进行信息资源的有效检索。

(2)WWW 的开放性允许网络用户可以任意地在 WWW 上发表自己的信息，这就导致了 WWW 上的信息质量的复杂性，使得 WWW 既包含大量的有用信息，也包含了大量无用的甚至有害的垃圾信息。

(3)由于包含了大量的垃圾信息，使得目前 WWW 的信息检索和 Web 数据挖掘的效率受到了很大的影响。在目前的检索系统中，用户无论全文检索还是根据关键字进行检索，都会发现在检索结果中有大量的信息与其查询目标相差甚远[1]。

Web 信息检索和挖掘技术能够帮助用户从海量信息资源中找到所需信息和隐含的知识，这在很大程度上缓解了信息和知识获取的困难。但是这些检索技术目前还存在以下问题：(1)返回给用户的检索结果中，有80%以上是用户所不感兴趣的内容，只有少量的信息能够满足用户的需求；(2)搜索引擎的搜索范围有限；(3)查询的方式有限，目前大多数的搜索引擎都是以关键字作为用户的查询接口。

WWW 上的信息挖掘(Web Mining)是 WWW 资源搜寻技术的一种，是数据挖掘(Data Mining)技术在信息发现技术领域中的应用。WWW 上的信息挖掘是指在目标样本的基础上，提取出目标数据对象间的内在模式，并以此为依据在 WWW 信息资源中进行有目的的信息提取。

但是，Web 信息检索和挖掘技术在目前远未达到成熟，还需要各个领域的研究人员做出更多的努力以实现高精度的检索和分类等工作。这不仅仅要求对检索、挖掘方法本身开展进一步的研究，同时还要在开展检索和挖掘工作之前对 Web 文档进行清洗等预处理工作。而后者在目前的绝大多数研究工作中往往被忽略。

各种检索和挖掘方法所取得的效果在很大程度上依赖所处理数据的质量。如果输入数据中包含了很多噪声或者坏样本，那么输出结果的可信度就会降低，甚至可能会出现错误的结果。由于多种原因，Web 文档不像传统的文本那样整齐、干净，其中包含了大量噪声，例如：为了增强用户交互性而加入的 Script，为了便于用户浏览而加入的导航链接，出于商业因素所加入的广告链接等；此外，与传统的文本文档相比，Web 文档在语义的内聚性上难以得到保证，有时一篇语义内聚的文章往往分散在若干个 Web 页面中，而有时一个 Web 页面中又包含了多个语义无关的部分。这些因素的存在，使得 Web 文档清洗工作具有特别重要的意义。如何有效地对 WWW 上的信息进行清洗，也就成为了 Web 技术领域中的一项重要研究课题。

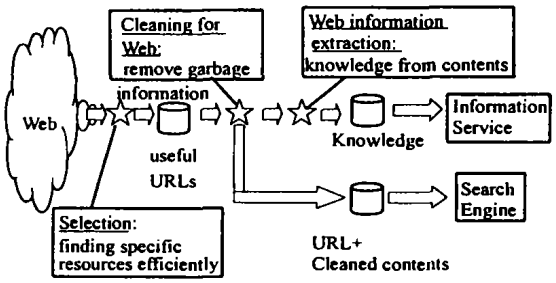

图1 Web 文档清洗的作用

*)本文得到国家自然科学基金(编号：600730307)和日本富士通研究所"Web 文档清洗"项目资助。张 波 硕士研究生，主要研究方向为 Web 信息检索与挖掘技术。王继成 博士，讲师，主要研究领域为 Web 信息检索与挖掘技术。王 强 硕士研究生，主要研究方向为多媒体文档处理技术。张福炎 教授，博士生导师，主要研究领域为多媒体技术、计算机图形学、数字化图书馆等。

Web 文档清洗的目标是根据检索、分类等后续应用的需求，将 Web 文档中杂乱的信息进行有效的整理归类，使之服务于不同的需求。Web 文档清洗所用的知识主要是机器学习、统计学、语言学等。经过 Web 清洗的文档，可以为信息抽取、信息检索、知识发现等后续应用服务（如图1）。

Web 文档清洗主要包含两个方面：(1)对图片信息的清洗。主要是将 Web 文档的各种图片根据后续应用的要求进行分类等，使得不同的后续应用可以进行不同的处理。(2)对文本信息的清洗。主要包括 Web 文档内容的合并与分割，文本块与链接块的区分等。

2 Web 文档清洗概念

2.1 什么是 Web 文档清洗

Web 文档清洗是根据后续不同应用的要求（比如 Web 信息检索、数据挖掘等）而对 Web 文档信息进行的过滤和筛选。不同的用户对清洗后的文档有不同的处理方法。例如，广告设计者可以仅对文档的广告图片进行检索，而网页设计者只需在各种装饰性图片中寻找他们满意的图案。因此清洗并不是将文档的内容删除，而是将文档的信息进行有效的整理分类，使得后续的应用能够根据不同的需求选取不同的内容进行处理。

Web 文档清洗的一般步骤如下：

(1)Web 文档的解析　是通过 HTML 解析器完成的。HTML 解析器将以文本方式保存的 HTML 源文件解析成 DOM 树。HTML 源文件中的各个元素则成为 DOM 树上的节点。

(2)特征提取　经过解析的 DOM 树中包含了相当大的数据量，系统为了有效地对其中的内容进行分类识别，就要对原始数据进行变换，得到最能反映分类本质的特征。这就是特征选择和提取的过程。一般我们把原始数据组成的空间叫测量空间，把分类识别赖以进行的空间叫特征空间，通过变换，可把在维数较高的测量空间中表示的模式变为在维数较低的特征空间中表示的模式。

(3)类型识别　就是在特征空间中把待识别对象（如图片、超链等）归为某一类别。基本作法是根据判决规则，对被识别对象进行分类。判决规则既可以手工制定，也可以通过机器学习获得。

(4)文档内容的输出　在识别出待分类对象后，对它们进行类型标识，获得比较“干净”的 Web 文档，并将这些文档输出为后续应用服务。后续应用对这些清洗过的 Web 文档进行信息检索、数据挖掘、知识发现等工作。

3 相关工作比较

清洗并不是 Web 所特有的一个课题。在数据库领域中，数据清洗（Data Cleaning）是数据挖掘（Data Mining）流程中的一个重要环节，其目标是过滤噪声，填补缺失的属性值，删除无效数据和重复记录[2]。应该说 Web 文档清洗在 Web 信息挖掘中所处的地位、作用与数据清洗在数据挖掘中相当，如图2所示。但是二者在方法上有着很大的不同。数据清洗处理的是结构化数据，通常可以在数据仓库的建立过程中完成；而 Web 文档清洗处理的是半结构化或无结构的 Web 文档，需要自然语言处理、机器学习等技术的辅助。

信息抽取是指根据预定义的模式从文档中识别出特定的部分。信息抽取的任务分为两类，一类着重利用语法规则来处理自然语言文本，另一类采用 Wrapper 从半结构化 Web 文档（通常由 CGI 等脚本语言从数据库中生成）中抽取信息。Web 文档清洗可以看作是信息抽取的一种特殊形式，可以借鉴信息抽取的一些研究成果。但是，Web 文档清洗需要处理的 Web 文档有时既不是自然语言文本，也不是从数据库中生成的具有良好结构的文档，因此还需要一些特殊的处理技术。

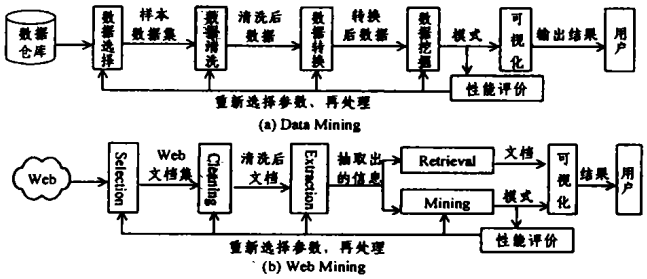

图2　Data Cleaning/Data Mining 与 Web Cleaning/Web Mining 的对比

目前，有一些系统能够从 Web 文档中去除广告，或者更为一般的，根据一些规范修改原始的 Web 文档，例如：Muffin，WebFilter，ByProxy 等。这些系统通常基于一些过滤规则来清洗 Web 文档。由于这些规则是手工编制的，因此只能够适用于特定的 Web 页面或者站点，无法随着 Web 的发展而扩展[3~5]。

4 总体框架

4.1 系统架构

图3给出了本系统的总体框架，其中包含两个部分：联机部分和脱机部分。

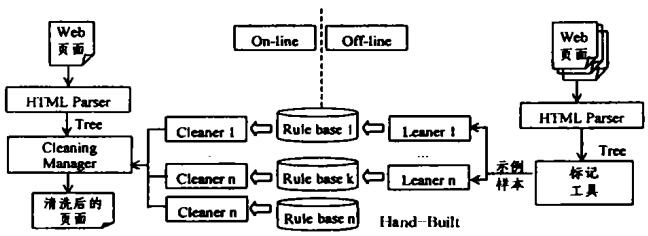

图3　Web 文档清洗系统总体框架

(1)脱机部分　其主要目的是构建用于 Web 清洗的规则库。该规则库中的规则用来决定 Web 文档的内容将要如何被清洗。近年来的许多研究成果表明，由于很多 Web 页面在表达内容时依靠的是语言学以外的一些因素，例如 HTML 标签等，因此，对 Web 页面的抽取和清洗并不一定需要复杂的语言学知识；基于一些简单模式建立的规则往往可以取得令人满意的效果。规则的构建通常可以依靠手工来完成。考虑到 Web 页面的庞大数量以及内容、格式的随意性，规则的构建有时也通过机器学习来实现。在本系统的实现中，将上述两种方法结合起来。对于图片信息的清洗和文本信息的清洗中的链接文字的区分，采用了机器学习的方法。对于文本信息的清

洗中的文本块和链接块的区分，则是采用了手工建立规则的方法。由于脱机部分主要是运行在后台，因此这部分对速度的要求并不高。

(2)联机部分　其目的是对于输入的每个 Web 页面进行联机清洗处理。具体而言，HTML 解析器首先对待处理的 Web 文档进行解析，得到其树状表示，然后清洗程序从中提取出各项特征并运用规则库中的清洗规则来抽取页面内容并进行类型区分和标识，最后将清洗结果输出给后续应用。联机部分由于直接和用户进行交互，因此要求有较快的速度。

4.2　规则获取步骤

我们先在收集到的各种样本上手工建立一些清洗规则。同时，采用机器学习的方法来自动生成另外的一些规则。从任务类型来看，主要包括以下几个方面的内容：①对于页面内容在粗粒度上的区分(文本块、链接块)将在 HTML 标签的基础上采用基于规则的方法来实现；②页面内容在细粒度上的区分(链接文字的区分与图片信息的区分)将在 HTML 标签、String Pattern、Logical Structure 等信息的基础上采用基于学习的方法。此时，用户不再手工构建规则，而是给出示例样本并进行正/反例的标记，由机器学习算法从中推导出相应的清洗规则[2]。以下介绍机器学习生成规则的步骤：

(1)样本的采集　采集的样本要具有一定的广泛性和代表性，其范围不能过于局限，否则，训练器生成的规则将不具有较好的推广性。理想条件是将 Internet 上的所有 Web 文档都作为样本，这种情况下的推广性将达到最好。但是，由于条件的限制，这种状态是不可能达到的。在实际的采集过程中，我们尽可能广泛地从不同类型的网站搜集样本。

(2)样本的标注　机器学习过程的开展必须建立在有效样本基础之上。样本的标注由人工完成。例如：为了学习广告图片的模式，必须先由人工在各种 Web 文档中找出广告图片和非广告图片，并获取相应的 HTML 代码，然后将这些代码样本提交给规则学习算法。这些工作的完成需要耗费大量的时间和精力，因此我们在 Web 文档清洗的后台开发了一些相应的标记工具，帮助用户完成样本的标注工作。用户只要点取 Web 页面上的图片或者超链接，并选择相应的类别，标注工具就可以在 HTML 代码中自动地插入标记。这样，我们就可以方便地对样本进行标注。

(3)HTML 文档的解析　由于 Web 文档的清洗需要建立在 DOM 树之上，因此 HTML 文档的解析是进行 Web 文档清洗的基础和前提。我们实现了一个 HTML 解析器，用于对 Web 文档进行解析。解析器生成的 DOM 树作为特征提取器的输入。

(4)清洗规则的建立　这一步的工作是由学习器负责完成。学习器的结构如图4。学习器在结构上主要由特征提取器和推导器组成：特征提取器从已生成的DOM树上提取出每幅图片的各种属性值，并将其进行处理，生成各种特征数据，并将其与预先定义好的特征模式一起作为推导器的输入，推导器最后生成规则库。

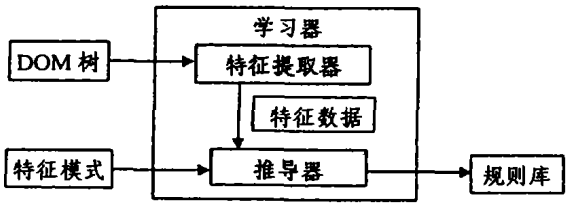

图4　学习器设计结构图

4.3　清洗步骤

(1)HTML 文档的解析　由于 Web 文档的清洗需要建立在 DOM 树之上，因此 HTML 文档的解析是进行 Web 文档清洗的基础和前提。我们实现一个 HTML 解析器，用于对 Web 文档进行解析。

(2)Web 内容的类型识别　利用后台学习和手工建立得到的清洗规则对新的 Web 文档进行清洗是一个相对简单的工作，主要是在 HTML 解析器得到的文档树的基础上，运用规则库从中区分出各种内容，并输出给用户。这一步由清洗器负责完成。

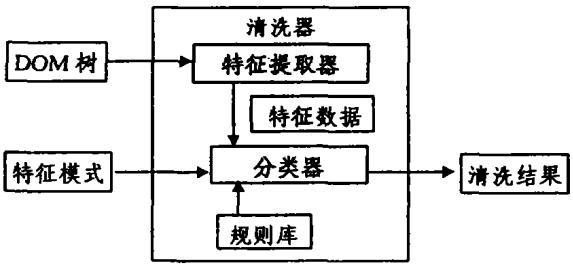

图5　清洗器设计结构图

清洗器在结构上由特征提取器和分类器组成(如图5)：特征提取器从已生成的 DOM 树上提取出欲分类的图片的各种属性值，并将其进行处理，生成各种特征值，并将其与预先定义好的特征模式一起作为分类器的输入，分类器根据已生成的规则库推导出最后的结果。

(3)Web 文档的清洗　根据上一步的结果，系统对各种类型的信息资源进行标识。在本文的研究中，标识是通过对 HTML 源文件进行修改进行的。清洗过的文档可以作为后续应用(如信息检索、数据挖掘、知识发现等)的输入。

结束语　随着 Internet 的迅速发展，Web 信息量急剧增长，Web 文档的清洗越来越显示出重要性和迫切性。本文提出了 Web 文档清洗的概念，详细介绍了 Web 文档清洗在 Web 信息处理过程中的作用及其步骤，并实现了一个基于规则的 Web 文档清洗系统。本文的研究成果被应用于南京大学与富士通公司合作的 Web 文档清洗项目中，取得了良好的效果。

参 考 文 献

1　王继成，张福炎，等. Web 信息检索研究进展. 计算机研究与发展，2001，38(2)：187～193

2　Fayyad U，et al. The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communications of the ACM，1996，39(11)：27～34

3　Kushmerick N. Learning to remove Internet advertisements. 3rd Int. Conf. on Autonomous Agents，1999

4　MUFFIN - World Wide Web Filtering System. http://muffin.doit.org/

5　Filtering the Web using WebFilter. http://math-www.uni-paderborn.de/~axel/NoShit/