文章编号：1003-0077(2009)05-0080-06

基于统计与正文特征的中文网页正文抽取研究

周佳颖[1,2]，朱珍民[1]，高晓芳[1,3]

（1. 中国科学院 计算技术研究所，北京 100190；
2. 湘潭大学 信息工程学院，湖南 湘潭 411105；
3. 首都师范大学 计算机科学联合研究院，北京 100037）

摘　要：该文提出了一种基于统计与正文特征的网页正文抽取方法。该方法继承了统计方法的优点，同时利用正文特征克服了原有基于统计的方法无法抽取多正文体网页的缺陷。源于多正文体在网页的 DOM 树中对应着正文区域下的多棵具有相似特征的正文子树，该文首先基于统计的方法获取一条正文路径，然后学习该路径的正文特征识别正文区域和子树主干，最后根据区域及该主干具有的正文特征进而得到完整的正文。实验表明该方法抽取单正文和多正文的精确率分别为 94％和 91％。

关键词：计算机应用；中文信息处理；正文抽取；单正文体；多正文体

中图分类号：TP391　　**文献标识码**：A

Research on Content Extraction from Chinese Web Page Based on Statistic and Content-Features

ZHOU Jiaying[1,2], ZHU Zhenmin[1], GAO Xiaofang[1,3]

(1. Institute of Computing Technology, The Chinese Academy of Sciences, Beijing 100190, China;
2. College of Information Technology, Xiangtan University, Xiangtan, Hunan 411105, China;
3. Joint Faculty of Computer Scientific Research, Capital Normal University, Beijing 100037, China)

Abstract: This paper presents a new method for content extraction from Web pages based on statistic and content-features. This method not only inherits the merits of the traditional statistic method, but also can extract the multi-body documents which can not be obtained by the pure statistic method. According to the fact that the multi-body documents are corresponding to multi-subtrees with the similar characteristics in the DOM tree of the web page, we first get a content path using the statistic method. Then, the content region and a trunk of subtree are modeled by the important features of the path, which are applied to get the whole information of the body content. Our experiment results show that the extraction precision of the single-body documents is 94％, and the multi-body documents is 91％.

Key words: computer application; Chinese information processing; content extraction; single-body documents; multi-body documents

1　引言

随着计算机和互联网技术的广泛应用，特别是随着“以人为本”的普适服务技术的发展[1]，如何有效地从种类和数量都急剧膨胀的网络中抽取出有价值的信息，为移动终端、个性化推荐等应用提供基础服务，是目前亟需解决的一个重要问题。正文抽取技术是一种广泛运用于互联网的数据挖掘技术。利用该技术分析巨大的网络信息源，并从中提取为咨

收稿日期：2008-09-04　**定稿日期**：2009-01-04

基金项目：国家“十一五”863 计划资助项目(2006AA01Z112)

作者简介：周佳颖(1983—)，女，硕士生，主要研究方向为普适计算；朱珍民(1962—)，男，博士，教授，主要研究方向为普适计算、嵌入式系统；高晓芳(1984—)，女，硕士生，主要研究方向为普适计算。

询、参考和决策提供支持的信息，以此构成普适服务的知识库。这是服务供应商提供可靠服务的基础环节。

然而，目前的正文抽取方法均有一定的不足。比如基于模板的方法受限于数据源[2-3]；基于视觉特征的方法实现复杂，扩展性不强[3]；基于统计的方法仅能抽取单正文体的网页。因此本文提出一种能克服上述三个问题的 SCF(Statistic and Content-Features)方法。该方法不限定数据源、易于实现、不需要人工参与、扩展了统计方法在多正文体网页的应用，并将其应用于医药建议系统的语料加工服务中。

2　相关工作

人们于 20 世纪 80 年代开始对 Web 信息抽取方法进行研究。目前常用方法主要有基于模板的方法、基于视觉特征的方法和基于统计的方法等三类。

基于模板的抽取方法是对特定数据源生成的网页集进行学习，不断发现、生成新的模板，从而建立模板库。文献[4]提出了利用网页链接分类算法和网页结构分离算法，抽取出信息并输出相应 Wrapper。文献[5]提出 Compute_CTokens 和 Construct_Template 两个算法来检测新的未知模板。但是这类方法限定于数据源。

网页中的标记信息既是网页的编写代码，同时也提供了页面的显示信息，利用这些信息，恰巧可以模拟人的视觉，因此出现了基于视觉特征的抽取方法。文献[6]最早提出模仿视觉特征来获取正文，文献[7]以 TABLE 标记为最小的容器单元和视觉特征来获取正文。然而，视觉特征非常复杂，不同的标签具有不同的规则，若要加入一条新规则，需要对原有标签的规则进行调整。

基于统计的方法通过统计各个标签包含的信息量来获取正文。此方法克服了限定数据源的缺点，具有一定的普遍性。文献[2]通过设定 2 个阈值 P，T 来过滤非正文的信息来获取正文。Mingqiu Song[3]不设定阈值，以 TABLE 节点作为统计信息的最小容器，利用公式 $WT = FC + 0.1 \times NC/HC$ 统计信息(其中 FC 为中文句号个数，HC 为超链接文字个数，NC 为非超链接文字个数，WT 为统计信息值)，最终得到一棵 WT 最大，仅有一个分支包含 TABLE 节点的 DOM 树，遍历树得到正文。但是该类方法，仅能处理正文内容在一个最小容器中的情况。

本文提出 SCF 方法来抽取网页正文。该方法巧妙利用正文特征的方法，克服了已有统计方法无法处理多正文体网页的缺陷。

3　SCF 正文抽取算法描述

本文提出了一个新颖的、更加有效的 SCF 方法，来自动挖掘网页中的数据记录(正文子树)。它能找到由表格包含的所有数据记录，表格标签包括 TABLE 和 DIV。它在不限定数据源、全自动化、实现简单的同时，能够处理单正文体网页(如新闻类网页)，或呈现在网页中是邻近且视觉效果相似的多正文体网页(如 bbs 网页)。该方法基于两个观测结果。

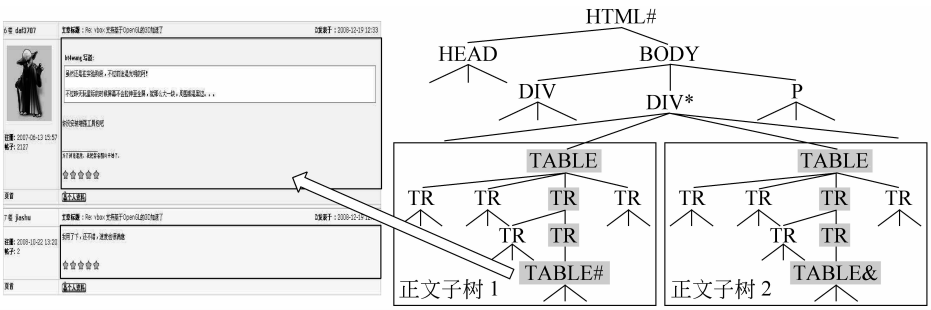


图 1　网页的片段记录及网页的标签树

1) 本方法处理的网页，仅有一个正文区域，该区域下包含一组相似的正文子树集，呈现在网页中是邻近且视觉效果类似的数据项，而且使用类似的 HTML 标签来布局。这样的区域为正文区域。如在图 1 中的两条讨论项，呈现在网页中是一个连续的区域，而在网页对应的标签树中，它们均位于一个正文区域(以 DIV * 为根节点的子树区域)下，使用几乎相同的 HTML 标签序列构造的一棵子树，如被矩形框围住的正文子树 1 和正文子树 2。

这种方法的问题是计算是受抑制的：基于统计

的方法，仅能在网页树中获取一条分支，根本无法直接识别一整棵子树。因此，需要找到一条能够代表该子树的分支，通过间接的方法来挖掘正文子树。下一个观测结果解决这个问题。

2) 正文子树均包含了一条具有相似特征的子树主干，呈现在网页中是一项数据记录的一部分，都是从子树的根节点开始，到底层容器节点终止的一条路径。任何两棵正文子树的子树主干，较长的主干中包含了短主干所具有的 HTML 标签序列，而且短主干的尾节点(底层容器)与长主干中的相应节点的部分属性相同(对应表格标签的视觉效果)。这样的主干为子树主干。如图 1，两条讨论项的子树主干，呈现在网页中是两个由矩形框包围的留言部分，而在标签树中，为正文子树 1 和正文子树 2 中节点底色加深的、从子树根节点开始到底层容器 TABLE 终止的路径，由于它们长度相同，因此均由一条〈TABLE，TR，TR，TABLE〉标签序列构造，且 TABLE# 与 TABLE& 节点的部分属性是相同的。

然而，本方法无法直接识别所有子树主干，因为多条位于正文区域下、代表正文子树的子树主干，不具有完全相同的嵌套层次。但是它们具有相似性，可以对构成它们的标签序列的相似度设定阈值。这样，只要识别出一条子树主干，便可以通过其具有的正文特征识别其他符合条件的子树主干。

据上述观察，得到正文子树(单正文体)的子树主干具有五项正文特征：A：子树主干能够代表正文子树；B：子树主干的父节点为正文区域；C：任意两条子树主干对应的标签序列的相似度大于某个阈值；D：短的子树主干的底层容器，与长子树主干的相应容器的部分属性相同；E：长的子树主干包含构造短子树主干的标签序列。

后续的实验证明这些观测结果是正确的。该方法通过四个步骤来实现：

第一步：构造网页标签树。

第二步：自动学习基于统计获取的一条正文路径包含的正文特征，来识别多条具有相同特征的路径，通过它们来识别正文区域及子树主干。

第三步：根据子树主干具有的正文特征，识别位于正文区域中的所有相似子树主干。

第四步：根据正文区域及子树主干集，剪枝并抽取。

本文创新点：

1) 关于“正文特征”这个要点，有从多正文体网页中的一个单正文体中，学习该网页正文体均具有的特征，并利用这些特征识别其他正文体。基于这种思想解决统计方法无法处理多正文体网页的缺陷。

2) 综合了“统计”和“正文特征”这两个要点后，整个系统不限定数据源、全自动化、实现简单的同时，能够处理单正文体网页(如新闻类网页)及多正文体网页(如 bbs 网页)。

3.1 标签树构造

将网页转换为 DOM 树，有两个环节：1)采用 NekoHTML 来转换不规范的 HTML 文档，并使用 NekoHTML 对网页进行解析，将 HTML 文档转化为 DOM 树。2)在 HTML 标签中，删除与抽取正文无关的节点，如文档的属性、显示风格、注释标记、图片、脚本。

3.2 挖掘正文区域及子树主干

本文间接挖掘正文区域及子树主干。首先基于统计的方法，识别一条具有信息量最大的、包含子树主干的正文路径，然后自动学习该路径包含的正文特征，去识别多条具有相同特征的路径，最后通过它们识别正文区域及子树主干。如图 1，首先基于统计的方法，假设第一条讨论项的信息量最大，则获取对应该条讨论的、从 HTML# 开始到 TABLE# 结尾的正文路径。学习该路径包含的特征：1)路径标签序列〈HTML，BODY，DIV，TABLE，TR，TR，TABLE〉；2)底层容器的属性集{class＝wr，cellSpacing＝0，cellPadding＝0，border＝0}。然后，根据该特征识别满足条件的其他路径，如从 HTML# 开始到 TABLE& 结尾的路径。最后通过它们识别正文区域 DIV＊及子树主干从 TABLE 开始到 TABLE# 结尾的一条路径。

定义 1 正文路径，是网页 DOM 树中一条包含正文的、从根节点到底层容器节点自顶向下的队列。

定义 2 子树主干，是正文子树中一条从子树根节点到底层容器节点自顶向下的队列。

定义 3 正文区域，包含了一条或多条，具有五项正文特征的子树主干。

这里主要分为两个阶段来描述。阶段一：通过统计的方法获取正文路径；阶段二：自动学习路径的特征并识别正文区域及子树主干。

3.2.1 获取正文路径

Mingqiu Song 方法能够获取一条最有可能包含正文的路径。本文在正文路径的获取上，优化了

他的方法：通过扩展统计信息的最小容器，使之能够处理多种类型的网页；根据制定的三个规则来清除非正文中句号 FC[3] 的影响，从而提高抽取正文的精确度。详细步骤如下：

1）扩展最小容器。将树中所有 DIV 节点替换为 TABLE 节点，DIV 节点在 DOM 树中的 4 类信息原样保留：与父亲的关系，与所有儿子的关系，节点具有的属性，原标签名。

2）根据规则识别装载非正文句号的容器并予以标记。经统计，非正文的句号主要出现在 3 处位置：(1)超链接的文字中；(2)网页版权信息中；(3)商家与用户进行互动的填表块中。在 DOM 树中，根据下述规则，识别并标记相应的容器。

规则 1：容器是包含链接地址的超链接容器；

规则 2：最底层 TABLE 容器标签的文本节点中出现了如“版权所有”或“Copyright”等类似语义的版权描述；

规则 3：容器为包含了要求输入、选择的控件和相关解释性文字的最小的容器。

3）基于标记统计 DOM 树中每个节点的信息：节点的 NC，HC，FC[3] 信息均为所有子节点中对应元素的和。其中如果节点被标记，则以该节点为根的子树的所有节点的 FC 置为 0。最终得到一棵信息 DOM 树。

4）采用 Mingqiu Song 方法，在信息树中获取一条具有信息量 WT[3] 最大的正文路径。

5）将原树转换得到的 TABLE 容器还原为 DIV。

3.2.2　挖掘正文区域及子树主干

由上述三条定义可以得出，正文路径＝网页树的根节点到正文区域的路径＋子树主干。根据这个等式能够挖掘正文区域及子树主干。本文首先自动学习初始获取的正文路径具有的特征，然后在网页树中，找到其他符合该特征的正文路径，最后识别覆盖所有正文路径最长的一条公共路径。公共路径的尾节点即正文区域，而在公共路径下开始分支的多条路径，则为子树主干。如图 1，在学习初始获取的正文路径的特征后，共识别了两条正文路径：从 HTML＃开始到 TABLE＃结尾的路径；从 HTML＃开始到 TABLE& 结尾的路径，它们的公共路径从 HTML＃开始到 DIV＊终止。这样便识别出正文区域 DIV＊，同时识别出两条子树主干：从 TABLE 开始到 TABLE＃结尾的路径；从 TABLE 开始到 TABLE& 结尾的路径。

下面具体描述算法的实现。

算法 1：自动学习路径的特征

输入：基于统计获取的初始正文路径

输出：路径标签序列，底层容器属性集。

1）从正文路径的起始节点开始到终止节点为止，依次读取节点的名称，并存入路径标签序列，

2）读取路径尾节点内的部分属性（包括属性名及其值），并保存到底层容器属性集，

3）终止。

此时已在线学习了正文路径包含的正文特征，其中第 2 步仅读取部分属性，因为实验中出现同是正文路径，但容器中一些属性存在差异的（如 id），因此这种属性，不予以保存。

根据正文特征信息：路径标签序列、底层容器属性集，遍历网页树，查找并存储所有符合该两个限制条件的正文路径。

算法 2：识别所有符合特征的正文路径

输入：DOM 树，路径标签序列，底层容器属性集

输出：正文路径集

1）将 DOM 树中，所有与路径标签序列中底层容器标签相同的节点，存入临时容器集，

2）从临时容器集中取出一个节点，

3）满足如下三个条件的路径为正文路径，存入正文路径集。(1)该节点为底层容器；(2)该节点的部分属性集与底层容器属性集完全相同；(3)从 DOM 树根节点，到该节点的标签序列与路径标签序列完全相同，

4）临时容器集不空，则跳到 2），

5）终止。

此时，根据正文路径集，来识别并存储正文区域及子树主干集，找出能覆盖该路径集中所有路径的、最长的一条公共路径。该路径的尾节点为正文区域。同时，位于正文区域下的、包含于正文路径中的子树主干也被识别。如图 1，从 HTML＃开始到 DIV＊结尾的路径，为覆盖对应正文子树 1 和 2 的正文路径的最长的公共路径。该公共路径的尾节点 DIV＊，就是要找的正文区域。而从 TABLE 开始到 TABLE＃结尾的路径，为对应正文子树 1 的子树主干。

3.3　挖掘相似子树主干

现在已经识别了正文区域及子树主干集，然而，之前的操作仅能识别完全符合正文特征的子树主干，可能一些具有相似正文特征的的主干还未被识别，因此还需经过两个步骤挖掘相似子树主干：

步骤一：找出所有候选主干。

根据下述两个条件找出正文区域下的所有候选主干，存入候选主干集。若候选主干集为空，则不用执行步骤二，否则还需要判别它们是否为子树主干。

候选主干满足如下两个条件：1)不存在于已获取子树主干代表的正文子树中；2)是一条从正文区域的孩子节点开始，到底层容器的路径。

步骤二：判别它们是否为子树主干。

取基于统计方法获取的初始正文路径所对应的子树主干作为代表主干(简称 d)，与位于正文区域下所有可能的候选主干(简称 c)，进行比较。依次满足下面三个条件的候选主干被认为是符合正文特征的主干，将其存入子树主干集。

1) $\min(|c|, |d|) > t \times \max(|c|, |d|)$(t 为设定的相似度阈值)；

2) $\min(|c|, |d|)$对应的主干的尾节点，与另一条主干相应节点的部分属性相等；

3) $\max(|c|, |d|)$对应的主干包含另一条主干的标签名序列。

如图 2，假设左边的主干为代表主干。右边的是候选主干，该主干满足上述条件：1)长度满足 $6 > (0.81 \times 7)$；2)主干中两个背景填灰的 DIV 节点的部分属性值相同，均为 class＝postbody；3)代表主干包含候选主干的标签序列〈TABLE，TR，TABLE，TR，TR，DIV〉。此候选主干符合正文特征，将其存入子树主干集。

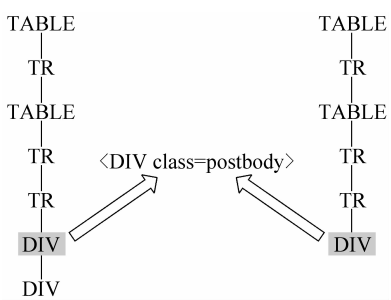


图 2 子树主干与候选主干

其中，条件 2)对应子树主干正文特征(见第 3 章节)的第 D 条特征；条件 3)对应第 E 条特征，由于字符串比较的时间复杂度高，因此将该条件放在最后；而条件 1)变相的对应第 C 条特征，变相是为了减少字符串的比较次数，下面详细解释条件 1)的来源。

本文基于 Levenshtein 距离[9]计算标签序列的距离。从而根据正文特征 E 推出子树主干 s1 与 s2 间的编辑距离为 $Edit(s1,s2) = \max(|s1|,|s2|) - \min(|s2|,|s2|)$。

本文采用非恒定的相似度评价系数(Jaccard 系数)[9]来描述两个字符串间的相异度，表示公式为：相异度＝r＋s/(q＋r＋s)，不难推出相似度公式为：相似度＝q(q＋r＋s)。其中 q 是 s1 和 s2 中存在的单词的总数，r 是 s1 中存在、s2 中不存在的单词总数，s 是 s2 中存在、s1 中不存在的单词总数。因此将两条子树主干对应的标签序列字符串及编辑距离代入相似度公式，得简化后标签序列相似度公式：$Sim(s1,s2) = \min(|s1|,|s2|)/\max(|s2|,|s2|)$。

本方法根据正文特征的第 C、E 条特征，将字符串的相似度大于阈值 t，转化为候选子树主干与代表主干间的长度关系，即为条件 1)。

3.4 剪枝与抽取方法

已有统计方法的剪枝，均在计算信息量的同时，就开始剪掉树中不满足条件的枝叶，那种方法仅适应于处理单正文体网页，无法处理多正文体的剪枝。本文的剪枝方法如下：

步骤一：在信息 DOM 树中，将标签树的根节点到正文区域的一条路径、所有子树主干中的根节点对应的子树两者包含的所有最小容器对应的节点做不剪切的标记。

步骤二：遍历树，若当前节点是最小容器节点且没有作不剪切的标记，则剪切该节点。

经过剪枝后，树中仅剩下包含正文部分的节点，遍历树得到的正文内容。

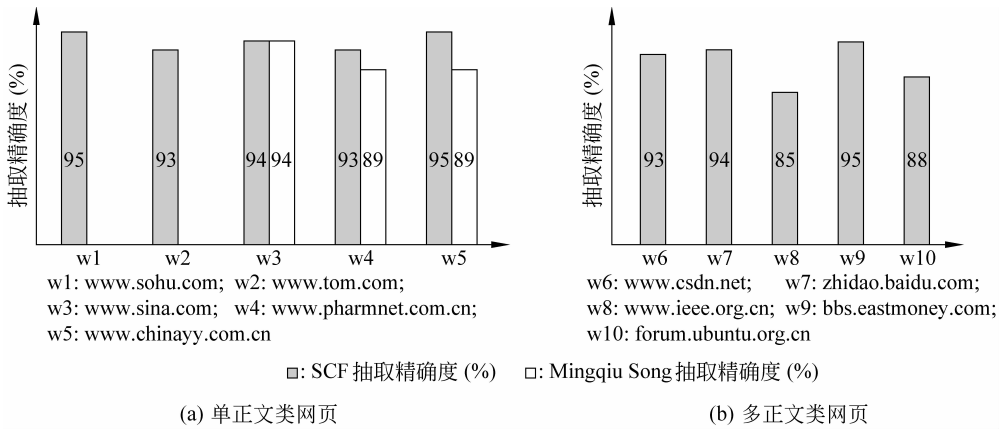
4 实验验证及结果分析

对一些实用性比较强的中文网站进行实验，主要挑选了单正文网页(新闻类、招商类网页，对应前 5 个网站)及多正文网页(bbs 网页，对应后 5 个网站)，进行连续 3 天的实验，每天在每个网站随机抽取 100 个网页，结果如图 3，在这个实验中，阈值 t＝0.81。

1) 单正文体网页

本方法与 Mingqiu Song 方法抽取单正文体网页的对比结果如图 3 中的坐标图(a)所示。

由坐标图可见，搜狐和 TOM 网站主要以 DIV 容器存储正文，而 Mingqiu Song 方法仅能处理以 TABLE 容器存储正文的网页，经过 3.2.1 节中扩展最小容器的处理，SCF 方法能抽取该类型的网页。对于后 2 个招商类型网站，SCF 方法抽取单正



(a) 单正文类网页　　(b) 多正文类网页

图 3　实验结果图

文的精确率比 Mingqiu Song 方法高 5%。这是在 3.2.2 节中，根据三条规则清除非正文中句号 *FC* 影响的结果。

2) 多正文体网页

SCF 方法抽取多正文体网页的结果如图 3 中的坐标图(b)所示。

考察出错的网页，发现大部分出错的原因：一是某些讨论项嵌套的引用别人的留言，该项对应的子树主干嵌套层次过多，造成该主干与其他子树主干对应的标签序列的相似度远远小于本文设定的阈值，最终造成抽取错误。二是基于统计的方法获取了一条错误的正文路径，从而导致提取了错误的正文特征，最终造成抽取错误。

从上面的实验结果可以看出，SCF 抽取网页正文，在不需要对同类网页学习，全自动化，实现简单，能抽取单、多正文体网页的同时，保持了较高的准确性。该方法抽取单正文及多正文的精确率分别为 94%和 91%。

5　总结与展望

SCF 方法抽取招商型单正体网页的精确率比 Mingqiu Song 方法高 5%；抽取多正文体网页的方法与基于模板的方法相比具有不限定数据源的优点，与基于视觉的方法相比具有实现简单的优点，与基于统计的方法相比具有能处理多正文体网页的优点。实际应用中，SCF 能处理呈现在网页中是邻近且视觉效果相似的多正文体网页。无法处理博客类网页，因为该类网页，博主发表的正文与读者评论的视觉效果不同，其内部结构也不同，下一步的工作是挖掘这种类型网页的正文。

参考文献：

[1] M Satyanarayanan. Pervasive Computing: Vision and Challenges[J]. IEEE Personal Communications, 2001, 6(8):10-17.

[2] 孙承杰，关毅. 基于统计的网页正文信息抽取方法的研究[J]. 中文信息学报, 2004,18(5):17-22.

[3] Mingqiu Song, Xintao Wu. Content Extraction from Web Pages Based on Chinese Punctuation Number [C]//Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007:5568-5570.

[4] 梅雪，程学旗，郭岩，等. 一种全自动生成网页信息抽取 Wrapper 的方法[J]. 中文信息学报，2008，22(1)：22-29.

[5] 杨少华，林海略，韩燕波. 针对模板生成网页的一种数据自动抽取方法[J]. 软件学报. 2008,19(2):209-223.

[6] Deng Cai, Yu Shipeng, Wen Jirong et al. VIPS: a vision-based page segmentation algorithm[R]. Microsoft Technical Report, MSR-TR-2003-79, 2003.

[7] 于满泉，陈铁睿，许洪波. 基于分块的网页信息解析器的研究与设计[J]. 计算机应用，2005，25(4)：974-976.

[8] Baeza-Yates, R. Algorithms for string matching: A survey.[J]. ACM SIGIR Forum, 1989, 23(3-4):34-58.

[9] 安淑芝. 数据仓库与数据挖掘[M]. 清华大学出版社，2005：115-119.