网页去重系统的设计与实现

樊勇，郑家恒，张虎

(1．山西大学 计算机与信息技术学院, 山西 太原 030006;

2．计算智能与中文信息处理省教育部重点实验室, 山西; 太原, 030006)

E-mail: sx-fanyong@163.com

摘 要:针对互连网上近似镜像网页严重降低搜索引擎效率的现象，本文提出了一种去重的方法。该方法通过句子在文本中的位置和组块的重要度，提取出网页正文的主题句向量，然后对主题句向量进行语义相似度计算，把重复的网页去除。实验证明，该方法在提高运算效率的同时，保证了较高的准确率和召回率。

关键词:组块；主题句向量；网页去重;

Design and Implementation of elimination of similar web page System

Fan Yong, Zheng Jia-heng, Zhang Hu

(1．Department of Computer and Information Technology, Shanxi University, Taiyuan 030006;

2．Key Laboratory of Education for Computation Intelligence and Chinese Information Processing，Taiyuan 030006)

E-mail: sx-fanyong@163.com

Abstract: Similar web pages that search engine returns not only waste storage resources but also increase the burden on web users. In this paper, a method to detect similar web pages is proposed. This method picks up topic sentence vector of web pages through location of the sentence in the text and importance of chunking. Then it detects the similar web pages by semantic calculate similar degree of topic sentence vector. The experiment results show that not only completely similar web pages are detected accurately but also partly similar web pages are detected exactly.

Key words: chunking; topic sentence vector; elimination of similar web pages

1 引言

随着互联网技术的飞速发展，网上信息在不断的爆炸性增长。根据CNNIC[1]发布的最新调查报告，截止到2007年6月30日，我国网站总数已经达到了131万个，而同时NETCRAFT[2]的报告显示全球的网站数量已经突破了一个亿。现有搜索引擎面临的最大一个问题就是返回结果集中包含大量的重复网页。这些网页有的是一字不差的完全重复，有的是其中一部分重复。大量重复网页的存在不但加重了用户浏览的负担，而且浪费了宝贵的存储资源、降低了索引效率，这也是影响搜索引擎质量的一个重要因素。因此，准确、快速的去除重复网页无疑是提高搜索引擎质量的关键技术之一。

现有的近似镜像网页去重方法大都来源于早期的文本复制检测算法。其主要思想就是按照一定规则从文本中抽取特征串来代替文本进行比较。特征串的抽取方法多种多样，Manber 提出的SIF[3]、Heintze 开发的 KOALA 系统[4]以及 Broder 等人提出的“shingling”方法[5]都是从文本中提

基金项目：本课题得到国家自然科学基金（60473139，60775041）项目资助。

作者简介：樊勇（1979-），男，硕士研究生，研究方向为自然语言处理。

取固定长度或固定个数的字符串作为文本特征。这些方法都是从网页的物理结构上机械的对其进行处理，对信息的内容没有进一步分析，缺乏语义支持。对于完全镜像网页的去重效果比较好，而对于近似镜像的网页去重效果不是十分理想。

本文提出了一种基于语义的网页去重方法，该方法分析了网页内容，对其内容进行主题的提取，并对主题进行语义相似度计算，将重复的网页去掉。实验证明，该方法不仅对镜像网页去重效果很好，而且对近似镜像网页去重效果也十分理想。

2 近似镜像网页正文结构的特点

本文中的18500篇网页是从Google和百度上搜索返回的结果中人工收集的，有散文类、新闻类的，对其分析具有以下特点：

(1) 结构松散。目前存储在介质上的Web 页面主要是用HTML 语言标记的文本文件，每个特定的Tag 对具有特定的意义，文本可以在页面的不同地方自由出现。Web 页面的文本字串缺乏普通文本中字串的属性，如位置信息(标题、小标题、段首和段尾等)。

(2) 主题杂合。Web 页面的内容和主题相对比较散，一个页面往往包含若干主题。为了显示其丰富的内容，许多网站都尽可能地在其页面，尤其是首页中放上各种内容和超链接，使页面中文本有时可多达1千到几千字。

(3) 处理不同。网页文本的内容处理也有异于传统文本处理。网络是一个虚拟的世界,包含了一种特殊的文化。比如，网虫喜欢把“东西”说成“东东”,将“漂亮妹妹”说成“PPMM”。这些词,在现代汉语常用词典里一般是查不到的。即使查到了,意思也大相径庭。因而传统的分词和匹配等处理方法在处理Web 页面时效果受影响。

3 系统主要架构

在本系统中使用的方法主要是利用了文本物理形式上的规律，如：组快的频率、标题、句子、段落等在文章中的位置等信息，结合《同义词词林》的语义知识，去除重复网页。结合了语义知识后，不仅对镜像网页去重效果很好，而且对近似镜像网页去重效果也十分理想。

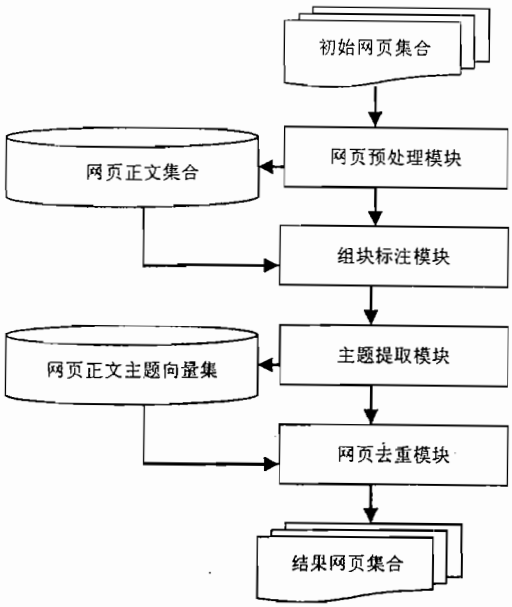
该网页去重系统主要分为下面几个步骤实现的：

1） 网页预处理，将导航信息、超链接信息、图片声音等信息屏蔽,获得网页的正文信息。

2） 组块标注，利用中国科学院计算所提供的汉语词法分析系统ICTCLAS对文本进行分词处理，然后按照词语在文中出现的顺序进行组块标注。

3） 提取主题，根据主题句在网页文本的位置给每一个句子加权，提取权重较大句子作为主题句向量。

4） 网页去重，对提取的主题向量集进行相似度计算，把相似度大于一定阈值的网页根据其内容多少和发布时间去重。系统结构如图所示：



系统结构图

4 系统实现主要步骤

4.1 网页预处理

本文使用了一种依靠统计信息，从中文网页中抽取正文内容的方法。首先根据网页中的HTML标记把网页表示成一棵树，然后利用树中每个结点包含的中文字符数从中选择包含正文信息的结点。这种方法克服了传统的网页内容抽取方法需要针对不同的数据源构造不同的包装器的缺点，具有简单、准确的特点。

4.2 组块标注

由于本文的实验系统对组块的划分质量要求不是很严格，因此，采用基于规则的方法对词语进行组块划分和处理。具体根据以下几个规则[6]完成：

（1）若有“和”“与”“跟”等连词，判断该词两边的词是否是同一种词性，如果词性相同，则合并成一个短语，短语词性与连词两边的词词性相同。

（2）若有多个名词相邻，将所有相邻的名词合并成为一个基本名词短语(BNP)。

（3）若有多个形容词相邻，则将其合并形成一个基本形容词短语(BADIJP)。

（4）将所有形容词跟其后面相邻的名词合并，作为基本名词短语(BNP)。

（5）若有“在”“于”等介词，后面是时间或处所以及与这些概念相关的名词，则将介词与后面的名词搭配形成基本处所或时间短语(BNP)。

（6）若数词之后的词为量词，则合并成一个基本数量词短语(BMP)。

（7）若有多个动词相邻，则合成为一个基本动词短语(BVP)。

（8）若有多个副词相邻，则合成为一个基本副词短语(BADVP)。

（9）若副词或形容词之后的词为动词，则跟动词合并成基本动词短语(BVP)。

4.3 主题句向量的生成

定义：一个主题就是一个“含义”，或者叫一个“概念”。它可以是一个词，也可以是一个短语，甚至是一个段落[7]。

对每个网页以句子作为一个研究对象进行分割。在对句子进行加权时我们不仅要考虑句子在文中的位置，还要考虑句子之间的关联信息。为了保留句子在文章中的表层信息，如位置、标题等信息，我们为句子这个对象定义了一种数据结构。具体数据结构如下：

```
StruSen{
int beginpoint;  //句子的起始位置
int senlength;  //句子的长度
ArrayList sencontent;  //句子内容
ArrayList splitsen   //句子分词后的结果
float weight;   //句子的权重
Boolean beginorend;   //句子是否为段首段尾句
String seqnum;   //句子在文章中的段号和句号
boolean title;   //句子在文章中是否为标题
boolean Stitle;   //句子在文章中是否为小标题
boolean linkv;   //句子结构是否为系表结构
boolean conclusion ;   //句子是否具有指示性
Wlist Chukwords;   //句子中包含的组快词列表
Boolean sentype;   //句子是否为陈述句
}
```

在对网页文本进行扫描分词时我们可以根据 StruSen 的结构对每一个句子进行初始化。beginpoint、senlength、sencontent、seqnum 根据实际情况赋值；weight初始值均设为0;beginorend、titile、linkv、conclusion 判断后相应赋值为 TRUE或 FALSE。

一般情况下主题句是陈述句，因此对于类型为陈述句的句子，根据以下加权公式计算该句子的权值[8]，句子类型不为陈述句时，将该句子的权值置为 0。

$$W(S_i) = A \times W_{be}(S_i) + B \times W_t(S_i) + C \times W_x(S_i) + D \times W_l(S_i) + E \times W_c(S_i) + F \times W_n(S_i)$$

其中 A.B,C,D，E，F 为比例系数，用来表明各因子在加权公式中的比重。具体权值如下表所示：

表1 主题句权重表

	句子特征	权值
$W_{be}(S_i)$	若 S_i 为段落的第一句	1
	若 S_i 为篇章的第一句	1.5
	若 S_i 段落的最后一句	0.6
	若 S_i 为篇章的最后一句	0.8
	其它	0
$W_t(S_i)$	若 S_i 包含标题关键词	1

	若 S_i 不包含标题关键词	0
$W_x(S_i)$	若 S_i 包含小标题关键词	1
	若 S_i 不包含小标题关键词	0
$W_l(S_i)$	若 S_i 为系表结构	1
	若 S_i 不为系表结构	0
$W_c(S_i)$	若 S_i 为指示性句子	1
	若 S_i 不为指示性句子	0

$W_n(S_i)$：表示句子 S_i 中包含组块的权值，具体取值如下：

$$W_n(S_i)=\frac{\sum_{i-1}^{n}F_{ij}\times W(C_i)}{M\times M'}$$

其中，M 为句子 S_i 包含的所有组块数，M′ 表示句子 S_i 中所包含的句数，$W(C_i)$ 表示组块 C_i 的重要度。

通过上面句子权重的计算可以得到一个主题句向量，但主题句的冗余度比较大。为了消除冗余，要进行主题句的相似度计算。降低冗余度后，形成一个初步的主题句向量，然后进一步加工形成最后的主题句向量。

4.4 网页去重

在网页去重这部分中，本文使用了基于语义的相似度计算方法，首先通过同义词词林对主题句进行编码，然后利用编码对主题句加权来计算相似度。具体公式如下：

A_i 和 B_j 的相似度计算：$$D(A_i,B_j)=\cos(T_{Ai*},T_{Bj*})=\frac{T_{Ai*}\bullet T_{Bj*}}{\|T_{Ai*}\|\times\|T_{Bj*}\|}$$

则 T_A 和 T_B 的相似度计算为：$$D(T_A,T_B)=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}D(A_i,B_j)}{n\times m}$$

其中 T_{Bj*}，T_{Ai*} 为主题句词权重向量。

为了计算方便对网页文本定义以下数据结构：

```
StruWwb{
ArrayList T_T;  //网页的标题
ArrayList T_t;  //网页小标题句向量
ArrayList T_s;  //网页主题句向量
}
```

为了提高去重算法的速度，本文采用了层次去重算法。具体方法如下：

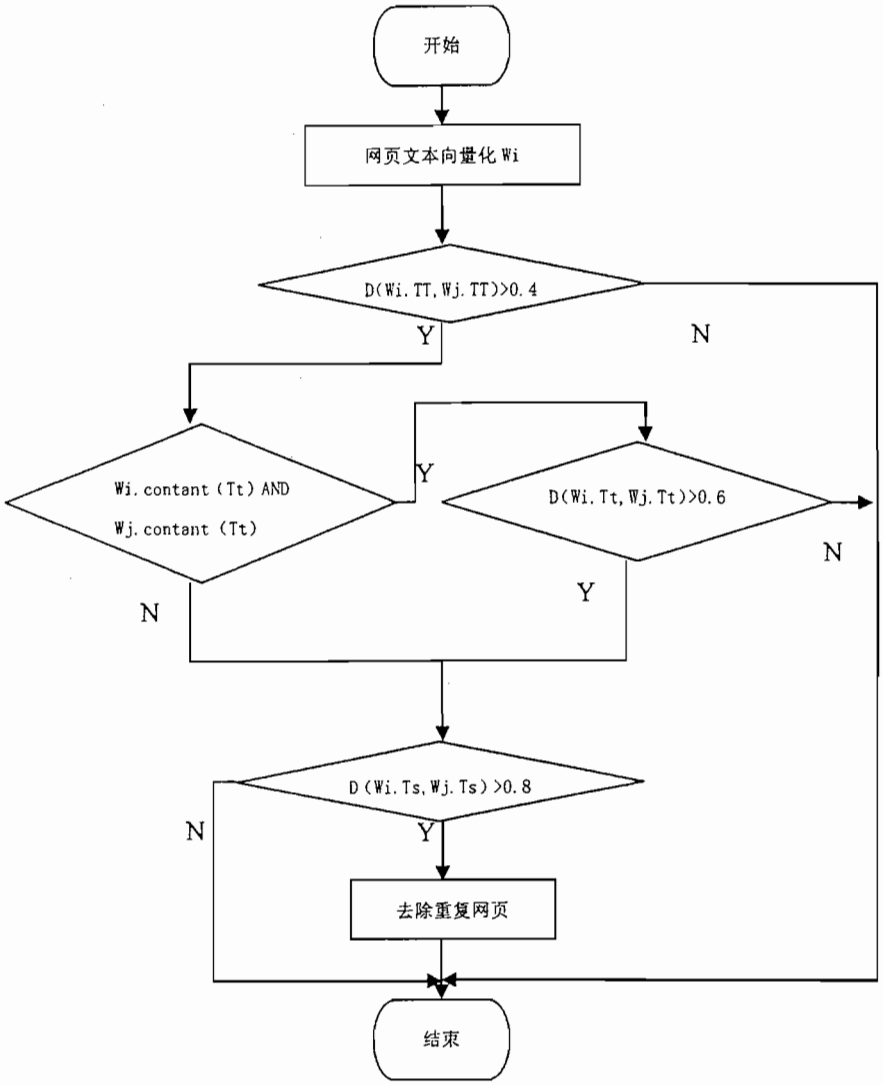
（1）对所有的网页文本进行向量化 W_i 其中 i 为网页文本总数。

（2）对标题句 T_T 进行相似度计算，由于互联网上同一篇文章可能采用不同的标题，所以标题相似为 0.4 以下的认为是不重复网页，否则转（3）。

（3）如果两篇文本都有小标题 T_t 的转（4），否则转（5）。

（4）计算其小标题句向量 T_t 的相似度，由于小标题能涵盖文本内容，因此小标题相似度为 0.6 以下的为不重复网页，否则转（5）。

（5）计算主题句向量 T_s 的相似度，其相似度为 0.8 以上的为重复网页，根据网页的发布时间和文本的长度去除发布时间较晚的或文本较短的网页。具体流程图如下：



5 系统测试结果

本实验所用的 18500 篇网页中散文类的有 2900 个，1119 个是部分重复的；新闻类的有 15600 个，6314 个是部分重复的。本文使用召回率、准确率和 F 值对算法性能进行评价，其中：

$$召回率=\frac{检测出正确的重复网页个数}{实际存在的重复网页个数}\times 100\%$$

$$准确率=\frac{检测出正确的重复网页个数}{检测出的重复网页个数}\times 100\% \qquad F值=\frac{2\times 召回率\times 准确率}{召回率+准确率}\times 100\%$$

实验一：散文类网页

	实际存在重复网页个数	检测出的重复网页个数	检测出正确的重复网页个数	准确率	召回率	F 值
全部重复	1781	1758	1696	96.58%	95.22%	95.91%
部分重复	1119	1090	984	90.28%	87.93%	89.09%

实验二：新闻类网页

	实际存在重复网页个数	检测出的重复网页个数	检测出正确的重复网页	准确率	召回率	F 值
全部重复	9286	9169	8973	97.86%	96.63%	97.24%
部分重复	6314	6065	5793	95.52%	91.75%	93.6%

对以上实验数据分析可以看出，基于语义的网页去重方法对于新闻类的网页效果很好，而对于散文类的网页去重效果一般，这是由于新闻类的网页结构化很强，散文类的没有固定结构。

本文还对系统的运行效率进行了测试，实验所用系统为：P4 3.2GHZ CPU， 1G RAM，完成时间是 468s。由此可见，本算法对新闻类网页不论是全文重复的还是部分重复的网页都有很好的检测效果，并且，快速的运行效率还使它能够适用于大规模的网页去重任务。

参 考 文 献

[1] 中国互联网信息中心. 第二十次中国互联网络发展状况统计报告[EB] http://www.cnnic.net.cn/index/0E/00/11/index.htm

[2] June 2007 Web Server Survey[EB]. http://news.netcraft.com/archives/2007/06/index.html

[3] Manber U. Finding similar files in a large file system. In: Proceedings of the Winter USENIX Conference. 1994. 1~10. ttp://manber.com/publications.html.

[4] Heintze N. Scalable document fingerprinting. In: Proceedings of the 2nd USENIX Workshop on Electronic Commerce. 1996. http://www.cs.cmu.edu/afs/cs/user/nch/www/koala/main.html.

[5] Broder AZ, Glassman SC, Manasse MS. Syntactic clustering of the Web. In: Proceedings of the 6th International Web Conference. 1997. http://gatekeeper.research.compaq.com/pub/DEC/SRC/technical-notes/SRC-1997-015-html/.

[6] 李素建, 刘群, 白硕. 统计和规则相结合的汉语组块分析. 计算机研究与发展, 2002年4月, 第39卷第4期:385-391. (期刊)

[7] 李盛韬. 基于主题的Web信息采集技术研究. 中国科学院, 2002届硕士学位论文. 41-43.

[8] 李立燕. 中文科技文献自动摘要系统。北京科技大学, 2006 届硕士学位论文. 35-37.