广西科学院学报
Journal of Guangxi Academy of Sciences

2009,25(4):251~253
Vol. 25,No. 4 November 2009

基于网页正文主题和摘要的网页去重算法*

The Detection on Duplicated Web Pages from Meta Search

周小平[1],黄家裕[2],刘连芳[1,2],梁一平[1],申文明[1]
ZHOU Xiao-ping[1], HUANG Jia-yu[2], LIU Lian-fang[1,2], LIANG Yi-ping[1], SHEN Wen-ming[1]

(1.广西大学计算机与电子信息学院,广西南宁 530004;2.南宁平方软件新技术有限公司,广西南宁 530003)
(1. School of Computer, Electronic and Information, Guangxi University, Nanning, Guangxi, 530004, China; 2. Pingsoft New Technology Co. Ltd. of Naning, Nanning, Guangxi, 530004, China)

摘要:针对元搜索返回的网页内容相同,别名差异很大的重复网页,提出基于网页正文主题和摘要的网页去重算法,并通过实验对算法进行有效性验证。该算法首先对各成员搜索引擎返回来的网页标题进行有关处理,提取出网页的主题信息,然后对摘要进行分词,再计算摘要的相似度,二者结合能更好地现出文章摘要的内容,实现网页去重。该算法有效,并且比基于传统特征码的算法有明显的优势,更接近人工统计结果。

关键词:去重 网页 分词 相似度 元搜索
中图法分类号:TP391.3 **文献标识码**:A **文章编号**:1002-7378(2009)04-0251-03

Abstract: According to the duplicated web pages returning from meta-search engine with same contents, but different name, an algorithm of duplicated webpages detection based on a combined duplication detection of the title and summary of web page is proposed. The effectiveness of the algorithm is verified through experiments. First, the algorithm analyze the page title which single search engines return; second, thematic information of page is extracted and word segmentation on the summary is carried out; finally, the similarity is calculated. By combining thematic information of web page title and the similarity of word segmentation on the summary, the algorithm can better to reflect the contents of the article summary, realize to detection and elimination of duplicated web pages. The algorithm has obvious advantages compared with the traditional signature-based algorithm, and is closer to artificial results.

Key words: duplicate detection, Web pages, Chinese word segmentation, repetition rate, meta search engine

最近几年来,由于科学技术的快速发展,网络得到了大量普及,互联网上的信息也呈"爆炸性"增长,用户要想查找自己想要的资料,就需要借助于搜索引擎的帮助。然而,各搜索引擎返回的结果中存在着大量的重复网页:(1)具有相同网址的简单重复网页;(2)具有相同别名和内容,但是由于转载造成了网址不同的名字相同的重复网页;(3)网页内容相同,别名差异很大的重复网页。这些重复的网页不仅浪费了存储空间,而且用户需要花大量的时间去查看这些重复网页,这就严重影响了查准率和浏览效率。尽量去除重复网页是搜索引擎关键技术之一。

网页去重是将搜集到的网页中的镜像或转载网页去掉的过程[1](镜像网页可以理解为一种特殊的转载网页)。几乎所有的网页去重技术都是基于这样一个基本思想:为每个文档计算出一组指纹,若两个

收稿日期:2009-10-10
作者简介:周小平(1981-),女,硕士研究生,主要从事信息检索和数据挖掘方面的研究。
*国家中小企业创新基金项目(编号:08c26224501313)资助。

文档拥有一定数量的相同指纹，则认为这两个文档的内容重叠性较高，也即二者是重复网页[1]。基于特征码的网页去重算法是将标点符号(比如“,”、“。”等等)作为分界点，分别依次从这些标点符号的两边各取 n 个汉字或字符作为该网页的特征码。清华大学采取在文章中出现逗号、句话的前后各取2个两字作为字符串，哈工大是在文章中每个句号的前后各取5个汉字[2]。虽然它们提取特征码采用的方法有差别，但是它们提取特征码的时候几乎都是以标点符号作为分界点，被提取的这些信息只代表了标点符号周围的信息，不能很好地体现出网页摘要的内容信息。元搜索引擎被称为搜索引擎之上的搜索引擎，它没有自己的Robot，只是提供某些接口去调用多个独立的搜索引擎(这些搜索引擎称为成员搜索引擎SSE)。当用户向元搜索引擎发出请求时，元搜索引擎就根据该请求向成员搜索引擎发出实际查询请求，然后将这些成员搜索引擎返回的结果进行去重、合并、重排序等处理，并以统一的格式返回给用户。元搜索引擎用到的网页去重技术是基于各成员搜索引擎对返回的摘要提取特征码，然后进行相似度计算来判断网页是否重复。

本文针对网页内容相同，别名差异很大的重复网页，提出基于网页正文主题和摘要的网页去重算法。该算法首先对各成员搜索引擎返回来的网页标题进行有关处理，提取出网页的主题信息，然后对摘要进行分词，再计算摘要的相似度，二者结合能更好地体现出文章摘要的内容。

1 基于网页正文主题和摘要

基于网页正文主题和摘要的去重算法主要思想是：提取两网页正文主题，判断是否相同或者相似，若是，则认为两网页是重复网页；否则，计算两摘要的相似度，判断是否是重复网页。

1.1 提取网页正文主题

算法首先处理返回结果中的网页标题。通过观察大量返回结果中的网页标题不难发现，网页标题完全相同的机率非常非常小，可以忽略不计。因此，问题的重点是在看似不相同的标题中找出相同或者相似信息。

网页标题不仅包含了网页正文主题信息，而且还包含了比如文章的来源等一些其它对去重无用的信息。信息之间通常用空格、“—”、“|”等隔开。首先我们借助分割符剔除无用信息，余下的就是能很好地体现出网页内容的主题。比如，在百度中输入查询词“java 截取字符串函数”，可得到一条标题：“Java面试中的一道编写一个截取字符串的函数－guofangsky的专栏”，利用分隔符“－”剔除无用信息“guofangsky的专栏”后，“Java面试中的一道编写一个截取字符串的函数”就是网页正文主题。

1.2 设计网页相似度评价函数

算法通过网页标题获取主题信息后，第2步是分别对主题信息和摘要进行分词，并计算各网页主题和摘要之间的相似度。

主题信息和摘要的分词利用JE-Analysis1.5.1中文分词器[3]进行，是采用字符串模糊匹配对分词后的主题和摘要进行相似度计算。假设有2个主题 t_1、t_2 和相应的摘要 S_1、S_2，分词后的主题分别记为 t_1'、t_2'，分词后的摘要分别记为 S_1' 和 $S'2$。

主题相似度评价函数定义为：

$\mathrm{TSimilarity}(T_1', T_2') = t_n / t_w\mathrm{Max} * 100\%$，

其中，t_n 为 t_1' 和 t_2' 中字符或汉字相同的个数；$t_w\mathrm{Max}$ 为 t_1' 和 t_2' 包含字符个数的较大值。

摘要相似度评价函数定义为：

$\mathrm{SSimilarity}(S_1', S_2') = s_n / s_w\mathrm{Max} * 100\%$，其中，$s_n$ 为 S_1' 和 S_2' 中字符或汉字相同的个数，$s_w\mathrm{Max}$ 为 S_1' 和 S_2' 包含字符个数的较大值。

1.3 网页相似度计算

算法描述如下。

初始化：输入某一查询词后，成员搜索引擎返回相关结果构成网页集合 $P = \{P_1, P_2, \cdots, P_m\}$，分别提取出每个网页的正文主题，构成主题集合，用 T 表示，$T = \{T_1, T_2, \cdots, T_m\}$；对主题进行分词，分词后的主题用 T' 表示，$T' = \{T_1', T_2', \cdots, {}_m'\}$，其中 m 为网页个数。然后对摘要进行分词，分词后的摘要用 S' 表示，$S' = \{S_1', S_2', \cdots, S_m'\}$，其中 m 为网页个数。

对于网页集合中的 W_i、W_j，判断两者是否是重复网页的步骤如下。

if (P_i. url=P_j. url)，P_i 跟 P_j 是重复的网页；

else if(T Similarity(T_i', , T_j')＞设定的阈值1)，P_i 跟 P_j 是重复的网页；

else if(SSimilarity(S_i', S_j')＞设定的阈值2)，P_i 跟 P_j 是重复的网页；

else if，P_i 跟 P_j 是不同的网页。

2 算法的实验及分析

2.1 去重效果比较

我们使用百度、谷歌作为元搜索引擎系统中的成员搜索引擎来进行实验。对于用户输入的某一查

询词,每个成员搜索引擎返回150条结果作为系统网页结果记录,存入数据库。分别输入6组具有代表性的关键词,对所得的结果分别进行人工分析和统计,将人工统计结果跟基于传统特征码的去重算法(简称算法1)和基于网页正文主题和摘要的网页去重算法(简称算法2)的去重结果进行比较,结果见表1。

表1 网页去重结果比较

组别	网页重复数(个)		
	人工统计	算法1	算法2
1	21	18	20
2	37	31	37
3	29	25	27
4	43	36	42
5	31	26	30
6	25	19	24

从表1可以看出,基于网页正文主题和摘要的网页去重算法(算法2)比基于传统特征码的算法(算法1)有明显的优势,更接近人工统计结果。

2.2 去重的正确率和召回率比较

使用谷歌和百度进行搜索得到500个网页,其中包括了86篇重复的网页,使用这500个网页进行实验。从表2可以看出,基于网页正文主题和摘要的网页去重算法(算法2)的正确率和召回率都明显高于基于传统特征码的算法(算法1)。

表2 各种算法正确率的比较

算法	实际存在的重复网页(个)	检测出的重复网页(个)	检测出正确的重复网页(个)	正确率(%)*	召回率(%)
算法1	86	75	69	92	80.2
算法2	86	83	81	97.5	94.1

*正确率=正确判重的网页个数÷判重网页总数×100%;召回率=正确判重的网页个数÷实际存在的重复网页个数×100%。

2.3 去重时间效率比较

对于输入查询词"eclipse java",百度和谷歌返回的结果分别为50条、100条、150条、200条、250条时,分别应用基于传统特征码的去重算法、基于网页正文主题和摘要的去重算法对数据库中的100条、200条、300条、400条、500条记录进行去重所花的时间如图1所示(实验中阈值1设为85%,阈值2设为70%)。从图1可以看出,基于网页正文主题和摘要的去重算法的去重时间要明显低于基于传统特征码的算法。

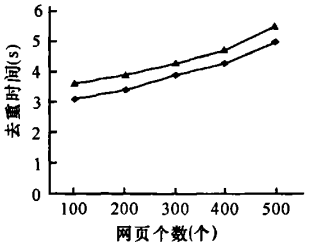

图1 两种去重算法的去重时间效率比较

-◆-:基于网页正文主题和摘要的去重算法;-▲-:基于传统特征码的去重算法。

3 结束语

在元搜索中,采用基于网页正文主题与摘要相结合的网页去重算法能有效地去除内容相同或相近的网页,提高网页检索质量,节约了用户查看资料的时间。未来网页去重的主要工作是通过大规模试验,进一步分析影响去重效果的因素,完善和改进算法,以期达到实用的目的。

参考文献:

[1] 李晓明,闫宏飞,王继民,等.搜索引擎——原理、技术与系统[M].北京:科学出版社,2004:112-115.

[2] 谢惠,秦杰.基于元搜索的网页消重方法研究[J].计算机系统应用,2008,17(8):94-96.

[3] 王学松.lucene+nutch 搜索引擎开发[M].北京:人民邮电出版社,2008:318-319.

[4] 谢惠,秦杰.基于用户查询关键词的网页去重方法研究[J].现代图书情报技术,2008(7):43-46.

[5] 刘迁,贾惠波.中文信息处理中自动分词技术的研究与展[J].计算机工程与应用,2006,42(3):175-177,182.

[6] 吴平博,陈群秀,马亮.基于特征串的大规模中文网页快速去重算法研究[J].中文信息学报,2003,17(2):28-35.

(责任编辑:邓大玉)