基于净化网页的改进消重算法①

虞 曼，熊前兴

(武汉理工大学 计算机科学与技术学院，武汉 430063)

摘 要：互联网的迅猛发展导致网络中的网页呈指数级别爆炸式增长。为解决在海量网页中寻找信息的问题，搜索引擎成为了人们使用互联网的重要工具。提出了一种基于净化网页的改进消重算法，并将它与传统的消重算法进行了比较。该算法结合关键字搜索和签名（计算指纹）搜索各自的优势来完成网页搜索消重。实验结果证明该方法对网页消重效果很好，提高了网页消重的查全率和查准率。

关键词：网页消重；净化网页；关键字；签名

Improved Duplicate Webpage's Elimination Algorithms Based on Purified Web Pages

YU Man, XIONG Qian-Xing

(College of Computer Science and Technology, Wuhan University of Technology, Wuhan 430063, China)

Abstract: The internet's development led to the rapid development on the explosive exponential growth level. To look for useful information, search engines have become one of the most important network tools. This paper presents an improved algorithm that is based on purified webpage and compared with the conventional algorithms. The algorithm combines the advantages of keyword search method and signature (calculated fingerprint) search method for the removal of duplicate pages. The experiments results certify that the algorithm improve the recall and precision.

Key words: duplicate webpage; elimination algorithm; Webpage purification; keywords; fingerprint

1 引言

随着互联网技术的飞速发展，互联提供给人们的信息资源越来越多，人们能够获得的信息资源也日益丰富。网络信息的指数级膨胀给信息检索带来了巨大的困难，并且随着人们使用互联网的越来越广泛，网络信息的易复制编辑使得互联网中存在大量的重复冗余的信息。

在信息检索过程中，这些大量的重复冗余信息对于用户而言，严重影响查询效率又掩盖了对用户真正有用的信息；而对检索系统而言，如果采集到大量重复网页，既浪费检索时间又浪费存储空间。网页消重技术的应用可以帮助用户快速去除查询网页集合中大量重复冗杂的结果以提高用户的查询效率。

2 网页消重技术的现状和前景

互联网中出现了越来越多内容重复甚至完全相同的网页（镜像网页），主题内容相同网页（转载网页）。就消除主题内容重复的网页而言，通常把镜像网页当做转载网页的特例来处理[1]。

近年来随着搜索引擎的越来越普及，针对网页消重的研究也越来越多的展开，目前代表性的方法有：基于模板消噪的方法、基于关键字匹配的方法、基于签名的方法、基于聚类的方法[2]。

(1) 基于模板消噪的方法

针对多数近似网页并非对原始网页的简单复制，而是采用类似内容置于新的模板之中。因此网页的模板会大大干扰算法对近似网页的计算。此方法是先对网页净化网页中的模板噪音内容，提取得到网页的正文，进而结合其他消重方法对其进行相似度计算消重。

(2) 基于关键词匹配的方法

此方法是利用向量空间模型[3](Vector Space Model

① 收稿时间:2011-04-08;收到修改稿时间:2011-05-22

—VSM)表示网页，并利用关键词进行相似度计算。首先对已抓取回来的网页集合 P 进行分析，提取网页中出现的关键词 T 作为网页的特征项。每个网页 pi 都分别提取相同个数的特征项关键词 tj，形成向量 Wi=(W1,W2,…Wj)。其中 Wj 的影响因素有两个：一是关键词 tj 在网页中出现的频率，二是网页集合 P 中出现关键词 j 的次数的倒数。而在判断两个网页是否为重复页面时，只需要判断表示两个页面的向量 Wi 和 Wj 的夹角的大小即可。

(3) 基于签名（即计算指纹）的方法

Narayanan Shivakumar 等提出了一种对全文分段匹配的算法[4]。这种算法是把一篇网页按一定的原则分成若干段（如 n 行作为一段或利用文本的自然段等），后对每一段进行签名（即计算指纹），于是一篇文档就可以用若干个签名后的指纹信息块来表示。U.manner[4]等提出一种对关键字签名的的方法。这种算法首先取网页的关键词按照频度排序，然后取频度较高的 N 个关键词进行签名运算。签名的结果作为表征网页的指纹。然后对签名进行相似比较作为相似网页判断的依据。

在这两个算法中，如果两个网页的 n 个指纹中有 m 个相同时（m 为定义的阀值），则可判定为重复网页。

3　算法实现描述

3.1 实现方式分析

常见的错误消重有以下两种情况：相同的内容但由于放在了不同的模板之中导致应该被消掉但实际上被消重程序误判定为非转载网页而保留；不同的内容但由于放在了相同的模板之中导致不应该被消重但却被消重程序误判定为转载网页而消掉。因此模板因素是导致消重不够准确的一个主要原因。为排除模板因素的干扰，使用仅包含网页标识、网页类型、内容类别、标题、关键词、摘要、正文、相关链接等要素的 DocView 模型[5]的正文代替网页原文参与消重。这样可以很好排除模板因素的干扰提高消重准确性。

针对关键词匹配的消重方法和基于签名消重方法各自的优缺点。本文给出了一种结合两种算法的方法。

3.2 算法描述

综合以上的分析，可以将算法简单的分为两个步骤来进行描述：

步骤一：采用有主题网页的信息提取算法提取网页集合之中网页的 DocView 模型。该算法以一组启发式规则为指导，先提取网页的正文信息，然后以此为基础提取 DocView 模型中其他的要素。下面按要素的生成顺序对其中的几个要素做简要描述：

正文：依据正文规则，深度优先遍历标签树并依次记录 topic 类型的内容块得到网页正文。

关键词：依据特征项的权值提取特征词。

内容类别：通过对正文类别分类得到的。

标题：提取网页标题<title>标签标识中的标题，或者选取权值最高的关键词作为网页标题。

摘要：基于关键词选择子句作为网页的摘要。

步骤二：对步骤一中净化后的网页采用消重算法进行消重处理。其算法可以用以下公式描述：

$$\left.\begin{array}{l}(MD5(Concatenat\ e(sort(Ti))) = \\ MD5(Concatenat\ e(sort(Ti)))\end{array}\right\} \Rightarrow Mirror(Pi,Pj)$$

$$SIM = \left(\frac{\sum_{a=1}^{N} W_{ia} * W_{ja}}{\sqrt{(\sum_{a=1}^{N} W_{ia})(\sum_{a=1}^{N} W_{ja})}}\right) < \delta$$

其中 $0 \le \delta \le 1$

在上述的算法描述中，

- Pi 表示第 i 个网页；
- Ti={t1,t2,t3,……,tin}表示网页权值最高的前 N 个关键字构成的特征项集合；
- Wi=<Wi1,Wi2,Wi3,.……Wim>表示网页权值最高的前 N 个关键字对应的特征向量；
- Concatenate（sort（Ti））表示对网页权值最高的 N 个关键词按字母序排序后拼接成的字符串；
- MD5（X）表示字符串 X 的 MD5 散列值；
- Mirror（Pi,Pj）表示 Pi 和 Pj 互为转载网页；
- A=>B 表示“若 A 成立则 B 成立”。

3.3 算法分析

由公式看出，设计的算法中条件一是表明如果两个网页的权值最高的前 N 个关键词集合使用 MD5 算法，判断其相同时就认为两个两页是相似网页需要进行消重处理，但不要求两个网页的权值最高的前 N 个关键词按权值顺序一致，这样能够避免了那些关键字权值不是顺序排列的情况的发生从而能够大大提高系统的查全率；但该算法的条件一只考虑网页的权值最高的前 N 个关键词，没有考虑这些特征项所构成向量的夹角的差异。在条件二中计算两个网页特征向量的相似度，使用夹角余弦值来定义度量两个夹角的大小。两个向量的夹角小两个网页则相似，反之则不相似。

在网页关键词的集合中，只提取了前 N 个权值较高的关键词。原因是特征值的前 N 个分量绝对值大，已经能够基本确定特征向量的方向。同时取较少的关

键词能够大大降低算法的复杂度。相似网页都是网页稍加改动，但基本意思没有变更。而网页的基本意思是通过其中出现的高频词来反映。特征集合中的那些没有考虑的低频特征词的出现时不稳定的。当使用这些词来判定相似网页会漏掉一大批相似网页。

在签名算法中本文对 N 个关键词组成的集合并没有要求相同。这是从算法的复杂度角度来考虑的，传统的签名算法需要判断两个集合的交集大小并求出交集。这个过程的时间复杂度很大，在海量所搜引擎中是无法实现的。因此，只考虑用 MD5 算法对关键词序列签名，来表示集合的相同与否。由于签名算法有极高的敏感性，作用对象稍有不同就会给结果带来很大的差异，并且无法从签名差异的大小来判断与原签名对象差异的大小[1]。因此简化条件一后可能会出现这样的情况，位置在 N 附近的词在排序上出现的稍小变动（如第 N 个词与第 N+1 个词位置更改），本来是两篇近似度很高的文章，可能被算法漏掉。因此这里增加条件二中 SIM 计算了两个网页向量的余弦值：

在搜索过程中，会遇到三种情况：当两个网页内容完全不相关时（即关键词集合没有交集），则 Wi 与 Wj 垂直，则 SIM 的值为 1；当两个网页的关键词集合相同且权值相同时，则 Wi 与 Wj 平行，则 SIM 的值为 0；当两个网页相似而不相同时，则 Wi 与 Wj 既不平行也不垂直，则 SIM 的值介于 0 和 1 之间。因此，SIM 的值成为判断两个网页相似度的标准。

4 算法评测

近似网页的算法的评价指标包括页查准率（Precision）、查全率（Recall）和算法复杂度（Algorithm Complexity）[6]。这些评测指标定义如下：

(1) 查准率，是指检测到正确的近似网页占总检测到的近似网页的百分比。假设算法检测到了 S 个近似网页，而其中有 So 个是正确的近似网页（即符合近似网页标准的网页），则算法的查准率为 Precision=（So/S）*100%。

(2) 查全率，是指检测到得近似网页占总的近似网页的百分比。假如算法检测到了 So 个正确的近似网页，而网页集合中实际存在了 Sn 个近似网页，则算法的查询率为 Recall=（So/Sn）*100%。

(3) 算法复杂度主要考虑算法的时间复杂度。将直接基于实验数据的计算得到。

还可以利用查准率和查全率的调和平均值[6]作为评测算法的综合指标。定义查全率为 r，查准率为 p，算法性能可以用 γ 表示：$\gamma = \frac{1}{\frac{1}{p} + \frac{1}{r}}$。

在实际的实验检测过程中是无法用大规模的数据库来进行的。需要在大数据的实验结果中进行采样，进行人工评测，去样本的准确率的平均值作为算法评测结果。算法运行的机器是一台 PC 机，配有 2CPU，内存为 2G，硬盘为 320G，运行的操作系统为 Windows XP。在这里随即选择规模为 10000 个网页作为实验的数据集。分别以基于关键字的消重算法、基于签名的消重算法及基于净化网页的改进算法进行测试并对比。其测试结果如下表所示：

表 1 三种算法评测结果对比（N=5， δ=0.1）

算法	总网页 S	近似网页 So	时间 T	p	r	γ
关键字	9887	9321	16M 43s	97.3%	97.0%	0.486
签名	9847	9532	17M 56s	98.2%	83.9%	0.452
改进算法	9867	9608	16M 17s	98.7%	95.7%	0.486

在表 1 中，可以看出基于净化网页的改进网页消重算法的查准率高于基于签名的消重算法和基于关键字的网页消重算法，而在查全率略低于基于关键字方法。但是在实际应用的过程，在庞大的网页数据库中要求的是更准确和更快的消重。在这个改进的算法中，N 和 δ 取值都大大会影响算法的查全率和查准率。而从实现算法的时间复杂度和空间复杂度上来看，算法的空间复杂比基于签名的网页消重方法要高，但是要远远低于基于聚类的网页消重算法，因此我认为在实际的应用过程中还是具有一定的应用价值。

参考文献

1 李晓明,阎宏飞,王继民.搜索引擎—原理、技术与系统.北京:科学出版杜,2005.95–115.

2 党春辉.网页消重和聚类算法在高校搜索引擎中的应用.上海:东华大学,2009.8–11.

3 杜海刚,李仙国.一种基于关键词的近似网页检测方法.微计算机应用,2008,2(2):42.

4 Shivakumar N. Finding Near-replicas of Documentson the Web. Proc. of Workshop on Web tabases. 1998:204–214.

5 张志刚.基于网页的信息系统的一种预处理过程.北京:北京大学,2010.23–25.

6 马文秀.近似镜像网页检测算法的研究及其评测.北京:北京大学,2006.25–28.