文章编号:1008-0570(2006)03-3-0113-03

基于特征码的网页去重

Duplicated Webpages Deletion Based on Feature Code

(桂林工学院)陈基漓 牛秦洲
Chen,Jili Niu,Qinzhou

摘要:网页去重处理是提高检索质量的有效途径,本文给出了一个基于特征码的网页去重算法,介绍了算法的具体实现步骤,采用二叉排序树实现。算法有较高的判断正确率,在信息检索中有较好的应用前景。
关键字:网页去重;网页特征码;二叉排序树
中图分类号:TP285 文献标识码:A

Abstract : Duplicated webpages deletion can improve quality of information retrieval. A duplicated webpages deletion algorithm based on feature code is given , the main steps of algorithm are introduced. the algorithm is realized on binary sort tree. The algorithm' s precision is high, has better application in information retrieval.
Keyword: Duplicated webpages deletion; feature code of webpages; binary sort tree

技术创新

1 引言

随着网络技术和信息技术的飞速发展,网络已经成为人们获取信息的一个重要途径。现有的搜索引擎面临的最大一个问题就是返回的结果集中包含大量重复的信息。如何更有效地帮助用户获取所需要的信息,能够快速、准确地为用户提供信息,是网络信息服务面临的新课题。优化搜索结果可以采用多种手段,如通过提取网页的特征进行基于内容的信息检索,利用用户反馈的信息进一步精确检索结果,将结果集中的重复信息尽可能地消除等。

由于网络信息分布的特点,网站上的信息存在相互转载及镜像站点等情况。出现相同网页主要有以下几种情形:网页的 URL 完全相同;网页的 URL 形式不同, 但网站域名所对应的 IP 是相同的;URL 虽然不同,但网页内容完全相同;URL 不同,为不同的网页形式,但网页上主要内容是相同的。本文主要讨论对于网页内容重复性的消除。

2 网页去重系统结构

网页为半结构化的信息形式,它与单纯的文本文档并不完全相同。网页中的有效信息主要包括以下几方面的内容:网页标题、网页正文、导航信息、超链接信息、图片声音等多媒体信息等。从以上信息中可以提取出有关网页内容的一些特征。

首先对检索结果集中的网页进行预处理,将其余信息屏蔽,获得网页的正文信息,然后用后面介绍的算法对网页正文进行去重处理。即判断是否已经有相同内容的网页出现在结果集中,若有,则进行删除或合并处理,若没有,则将该网页保留在检索结果集中。网页去重系统主要结构如图 1 所示。

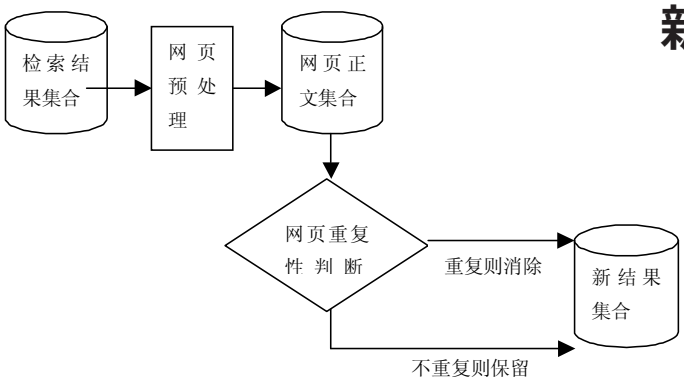

图 1 网页去重系统结构

3 基于特征码的网页去重算法

对网页进行去重处理,实质上是从一批网页中将内容相同或相近的网页分为一类,进行聚类处理。用传统算法进行聚类处理,只能将同一大类的网页聚合为一类,与传统意义上的聚类处理不同,网页去重需要对网页进行较为精确的归类。如果严格按照网页内容进行分类,则分类结果中类别会很大,导致在确定一个网页属于哪一类时计算所花费的时间过大。如果直接将网页正文逐字进行匹配处理来实现归类,也同样会出现计算量过大,而在响应时间上无法承受的问题。较好的方法是从网页正文中抽取出少量信息构成特征码,在归类时,以特征码取代网页,通过判断特征码是否相同或相近来判断相应的网页内容是否是重复的。

3.1 网页特征码

陈基漓:讲师
基金项目:广西区科技攻关项目 (桂科攻 0428002-1)

网页特征码首先必须能够较为全面地反映网页的内容,其次为了计算上的方便,特征码在长度上有一定的限制,不能太长。

采用以下方法构造网页特征码:特征码由主码和辅码两部分构成。依次提取网页正文中每段段首的第一个字,组成主码。再从各段中将每一个标点符号前面的一个字提取出来,依次构成辅码,考虑特征码长度方面的限制,辅码提取中只对每段的前 n 个标点符号进行提取。若某一网页正文共有 5 段,取 n 值为 3,则提取出来的网页特征码结构如图 2 所示。

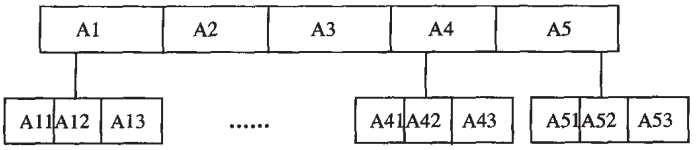

图 2　特征码结构

3.2 网页重复性判断算法

提取出网页的特征码之后,下一步工作是依据特征码判断网页正文是否重复。假设网页 a 对应的特征码为 Ta,网页 b 所对应的特征码为 Tb,判断 a 与 b 是否为重复网页的主要步骤为:

1)比较 Ta 与 Tb 的主码部分,若两者主码完全相同,则认为网页 a 与网页 b 是内容相同的网页,转 4),否则转 2);

2)若 Ta 与 Tb 的主码比较结果为以下情形之一:①其中一个的主码为另一个主码的真子集;②两者不互为真子集,但两者主码取交集的结果较大,则转 3)作进一步判断;若两者主码取交集为空或交集结果较小,则认为网页 a 与网页 b 是内容不同的网页,转 4);

3)对 Ta、Tb 主码的交集,即两者相同的主码部分进行处理,判断对应的辅码是否相同,若完全相同或大部分相同,则认为网页 a 与网页 b 是内容相同的网页,若相同的辅码很少,则认为网页 a 与网页 b 是不同的网页,转 4);

4)算法结束。

在判断算法中,对于以下情况认为两个网页是相同的:一个网页内容是另一个网页的部分内容,或两个网页虽然不完全相同,但其中大部分内容是相同的。可以通过设定一定的阈值对算法中的不确定因素进行判定。如两者交集结果超过其中任何一个的 80%,则表示两者交集结果较大,反之当小于 20%时,认为两者交集结果较小;在对辅码进行比较时,当相同的辅码占 80%以上时,认为辅码大部分相同。可以根据实际检索的结果,将阈值调整至一个比较合适的取值范围,获取较为满意的检索结果。

3.3 算法有效性分析

网页重复性判断算法是否有效,关键是特征码与网页正文内容之间的对应关系,若不同内容的网页对应的特征码是不同的,则保证了算法的有效性。若出现多个不同内容的网页有相同的特征码,则会将不同内容的网页归并到一类进行处理。若单纯从文字上看,以中文网页为例,常用的汉字大约为 6700 个,特征码主码的长度为 n,则对于不同网页出现相同特征码主码的概率为 $1/(6700)^n$。虽然对于一些热门新闻,段首文字多以‘据报道”、“新华社’等文字开头,若有 m 段文字以这样的固定词开头,(m 小于 n),出现重复特征码的概率为 $1/(6700)^{n-m}$,当 n-m 或 n 的取值稍大,如大于 5 时,这样的概率值是很小的。同时在算法中,还考虑了辅码的作用,当出现主码部分相同时,进一步判断辅码的分布以确定特征码是否相同。

4　算法实现

4.1 数据结构的选择

检索结果集中的网页具有动态变化和数量巨大两个特征,必须选择一种合适的数据结构,减少去重过程(相同网页合并过程)的比较次数,同时又能较好地表示动态变化的特征码集合。

二叉排序树能较好地满足上述要求,选择二叉排序树作为算法实现的主要数据结构。二叉排序树或为一空树;或是具有下列特征的二叉树:1)若左子树不空,则左子树上所有结点的值均小于它的根结点的值;2)若右子树不空,则右子树上所有结点的值均大于它的根结点的值;3)它的左、右子树也分别为二叉排序树。

为描述方便,以下所说相等是指两个特征码对应的网页内容相同或相近,不等是指两个特征码对应的网页内容不同。当出现特征码相等的情况时,需要进行合并处理。算法采用扩展二叉排序树为主要的数据结构,在传统的二叉排序树的每个结点中增加一个指针,该指针指向由特征码构成的链表,称为‘辅指针”。辅指针指向与该结点对应网页内容相近的网页特征码,这样就可以较方便地处理网页合并的情形。最终保留在检索结果集中的是扩展二叉排序树中各结点表示的特征码对应的网页。扩展二叉排序树结构如图 3 所示。

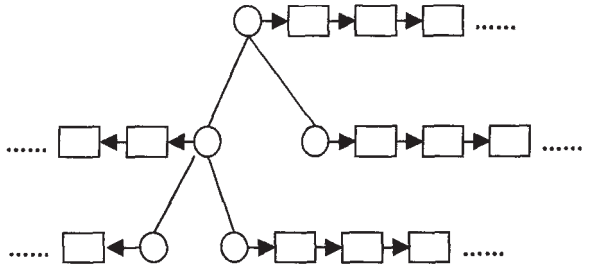

图 3　扩展二叉排序树结构

4.2 特征码归类过程

二叉排序树的构建过程也就是对特征码进行归类处理的过程。对于一个新处理的特征码,在二叉排序树中没有找到可以合并的结点时,直接对该特征码

技术创新

进行插入操作即可。在二叉排序树中找到相等的特征码时，该特征码要进行合并操作，而不同于普通意义上二叉排序树的查找操作。

当二叉排序树为空时，将新处理的特征码作为新结点插入树中，插入新结点时，该结点的辅指针为空。当二叉排序树非空时，首先将新处理的特征码与根结点表示的特征码比较，若相等，则进行合并处理，若不等，则根据新处理的特征码与根结点表示的特征码之间的大小关系，分别在左子树或右子树上继续进行比较，在比较过程中，若出现相等的情形，则将新处理的特征码与相应的结点进行合并，若在整个比较过程中，始终出现的是不等的情况，则说明新处理的特征码所对应的网页内容还没有出现在二叉排序树中，将其作为一个新接点插入。

假设网页 x 对应的特征码为 Tx，网页 y 所对应的特征码为 Ty，Tx 为新处理的特征码，Ty 为二叉排序树中出现的与 Tx 相等的特征码，采取以下策略进行合并：1) 若 Tx 与 Ty 的主码完全相同，则二叉排序树不需要做任何改动，直接将检索结果集中网页 x 删除；2) 若 Tx 主码为 Ty 主码的真子集，则将 Tx 与 Ty 辅指针所指向链表中各结点的特征码进行比较，若无相等的特征码，则将 Tx 作为一个新结点插入 Ty 辅指针所指向链表中；3) 若 Ty 主码为 Tx 主码的真子集，将 Tx 取代二叉排序树中结点 Ty，同时将 Ty 依据 2) 中的原则插入 Tx 辅指针所指向链表中；4) Tx 与 Ty 不互为真子集，但两者主码取交集的结果较大，处理方法同 2)。

4.3 算法效率分析

不管新处理特征码是进行合并或插入，均要先进行查找比较，已确定插入的位置或合并的结点。对二叉排序树进行比较，在结点出现概率为随机概率分布的情况下，平均查找长度小于等于 $2(1+1/m)\ln m$，m 为二叉排序树中结点的个数，平均查找长度与 $\log m$ 成数量级，即比较过程的时间复杂度为 $O(\log m)$。插入结点过程只是一些指针的移动，时间可以忽略不计。

由于特征码主要由各段段首字及每段中前 n 个标点符号前的文字构成，因此对特征码的提取不需要对整个网页正文都扫描一次，特征码的提取时间与处理的网页正文长度有关，可以看成是一线性关系，特征码提取的时间复杂度为 O(n)。

4.4 聚集处理

为进一步提高算法的效率，将特征码进行聚集处理，即将具有某种相同特性的特征码放在同一棵扩展二叉排序树中，整个算法采用由若干棵二叉排序树构成的森林。在处理一个特征码时，首先判断应该属于哪棵二叉树，然后再根据合并算法进行处理。

以正文的长度（字节数）tl 和特征码中主码的长度 kl 作为聚集的主要衡量因素，取聚集因子 $\alpha=tl/kl$，将 α 分成若干个区间，如 (0, 50], (50, 100], (150, 200].....，根据特征码对应的聚集因子确定其所属的二叉树。经过聚集处理后，处理一个特征码时，只需要与和它具有类似聚集因子的特征码进行比较，比较的次数将在一定程度上减少。

5 实验结果

采用以上介绍的算法，对一批数量在 50-100 之间的网页集合进行处理，集合中包含了一些内容完全相同或部分相同的网页，将实验结果与人工判别的结果进行比较，发现重复网页的正确率达到 95%以上，出现错误的判断有些是由于网页转载时出现错码等现象，有的是两个重复网页的段落排列差异太大。

对算法的执行时间做测试，测试结果如下表所示。从实验结果可以看出，去重处理过程中主要时间用于特征码的提取。

网页数量（个）	特征码提取时间(s)	去重处理总时间(s)
500	3	4
1000	7	9
1500	10	13
2000	15	18
3000	26	31
4000	35	41
5000	43	50

技术创新

6 结束语

基于特征码的网页去重算法，能有效地去除检索结果集合中内容相同或相近的网页，将这项技术用于网络信息检索系统中，可以提高检索质量，使返回给用户的结果更为精确，有较好的实际应用前景。未来的工作主要时两个方面，一是进一步完善算法，在算法中加入内容的判断，以期解决由于段落排列差异大而造成的判断错误；其次是采用更好的方法进行聚集处理，使聚集的结果更合理，更符合网页本身的特征，以适应对更大规模网页集合的去重处理。

参考文献：

[1]谢立,王永强,于德敏,许增朴. 利用图像的灰度特征实现半透明产品的识别[J],微计算机信息, 2005,7:44-46.

[2]Finding near-replicas of documents on the web. Narayanan Shivakumar, et al. WebDB 1998

[3]Finding replicated web collections. Junghoo Cho, N. Shivakumar et al. In Proceedings of 2000 ACM International Conference on Management of Data (SIGMOD), May 2000.

[4]数据结构(C语言版),严蔚敏,吴伟民.清华大学出版社,1997.

作者信息：陈基漓(1972—)，女，硕士，讲师，主要研究方向：信息检索，数据库，e-mail: zhchenjili@sohu.com；牛秦洲,男,博士,教授,主要研究方向：计算机网络控制。(541004 广西桂林 桂林工学院电子与计算机系)陈基漓 牛秦洲

(Depatment of Electronics and Computer,Guilin Institute of Technology, Guangxi, Guilin 541004) Chen,Jili Niu,Qinzhou

(投稿日期:2005.8.1) (修稿日期:2005.8.8)