MD5 算法在农业数据消重中的应用①

Application of MD5 in Agricultural Data Cleaning

刘　峰[1,2]　王儒敬[1]　(1 中国科学院 合肥智能机械研究所 安徽 合肥 230031)
(2 中国科学技术大学 自动化系 安徽 合肥 230027)

摘　要： 垂直搜索引擎的数据来源于各大相关网站。随着互联网用户越来越多，相互转载，重复发布的数据也越来越多，由于这些重复及相似数据的存在，严重影响了搜索引擎的检索结果。针对这一问题本文提出了一种解决方法，即利用 MD5 算法在数据处理阶段有效的消除冗余数据。结果表明该方法很好的提高了检索质量。

关键词： 垂直搜索引擎 检索结果 MD5 算法 冗余数据

1　引言

随着互联网的不断发展，人们越来越多地在互联网上发布和获取信息。Web 已经成为信息制造、发布、加工和处理的主要平台。传统的互联网应用技术大多是基于文档内容的，与经典的信息检索技术和数据库技术有着密切的联系。但是，互联网中特有的许多问题，使得互联网应用技术很难有效地应用。大量重复的网页数据就是其中的一个问题：有 41 % 的网页是具有 50 % 的相似性[1]。Stanford 的 Cho 等人在 1999 年利用 Google 搜索到的 25,000,000 个网页的数据集统计得出约 48 % 的网页是重复的[2]。而垂直搜索引擎面向领域的特点使得这种现象更加突出，根据我们采集到的农业数据进行的统计，每天采集到的数据有 60%左右是重复的。在这种背景下，本文提出了使用 MD5 算法对数据进行消重，在数据处理阶段进行有效的冗余数据消重。

2　算法思想及算法描述

2.1 算法思想

MD5 的全称是 Message-digest Algorithm 5（信息-摘要算法），在 90 年代初由 MIT Laboratory for Computer Science 和 RSA Data Security Inc, 的 Ronald.L.Rivest 开发出来。它的作用是让大容量信息在用数字签名软件签署私人密钥前被"压缩"成一种保密的格式（就是把一个任意长度的字节串变换成一定长的大整数）。对 MD5 算法简要的叙述为：MD5 以 512 位分组来处理输入的信息，且每一分组又被划分为 16 个 32 位子分组，经过了一系列的处理后，算法的输出由四个 32 位分组组成，将这四个 32 位分组级联后将生成一个 128 位散列值[3]。本文即是使用 MD5 算法将农业供求数据不同的字段值经过处理得到一个 128 位的 MD5 编码存放于数据库的表中，并对该表建立索引，然后通过比较 MD5 编码迅速消除数据库中重复及相似的供求数据。

北京大学的天网使用了多重 MD5 算法，其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，也即两者是互为近似相似的[4]。

假设数据集为 N 个网页，若我们为每个网页生成 m 个指纹，则对于 N 个网页两两比较以检测近似的时

① 基金项目:国家 863 计划(2006AA10Z23702);国家科技支撑计划(2006BAD10A0502);国家科技支撑计划(2006BAD10A1410);国家自然科学基金(60774096)
收稿时间:2008-07-18

间代价就是 O (m^2N^2)。这样的计算量，对于 N 为上亿个 Web 页面的数据集来说是不可能完成的。

所以，一般的全文签名算法都会对数据作一些处理以降低算法的复杂度。有算法使用对<文档标识(DocID),段标识(ChunkID),指纹(Fingerprint)>三元组排序的方法避免了对所有网页作两两比较[5]，使算法复杂度有所降低。但是该算法的空间复杂度和时间复杂度仍然是相当大的，若应用于海量的搜索引擎系统，仍然难以取得理想的效果。

天网的算法在生成每个网页的指纹的时候，对于这 m 个指纹进行了优先级排序（例如可按照所对应文本块的大小）。这样，在比较网页指纹组的时候，只对于在前 pre（0<= pre<= m）个位置上存在相同指纹的网页对进行两两比较。由于前几个指纹多数对应的是网页文本的标题块和主要正文块。该算法可以大大降低算法的复杂度，但是对于农业垂直搜索引擎要每天及时更新数据库来说，这也难以满足要求。

在考虑到农业垂直搜索引擎数据的特点后，发现农业数据完全重复的居多，所以本文采用单 MD5 指纹对完全相同的数据进行消重后，再对消重后的数据利用多重 MD5 指纹消重，且定义 pre=3，因为在本文设计的农业垂直搜索引擎中使用的文档数据前三个指纹对应的是文档标题，联系人以及联系方式，只有这三个字段一样的情况下我们才认为供求信息可能是相似的。

2.2 算法描述

根据前文的描述，本文提出了单 MD5 和多重 MD5 混合使用的算法，其算法流程如下：

第一步：单 MD5 指纹消重

①从数据库供求信息表（即存放原始供求信息数据的表）中读取供求数据各个字段值并连接生成字符串；

②对 1 中生成的字符串进行 MD5 编码，得到编码值，并写入数据库供求信息临时表中（该表比供求信息表多了一个字段即 MD5 编码字段）；

③对数据库中供求信息临时表建立索引；

④按照时间段在索引文件中进行 RangeQuery 查询产生结果集 hits1；

⑤取出 hits1(i)的第一条记录 hits1(0)，在索引文件中进行 TermQuery 查询生生结果集 hist2；

⑥如果 hist2.length()>1，说明有重复记录，得到重复记录的 ID，在数据中删除，并在索引文件中同步删除；

⑦将步骤 6 中重复记录中 REG_TIME 值最大的记录写入数据库以及索引文件中；

⑧i++，返回步骤 5 中取下一条记录，直到 hits1(i)记录为空结束。

第二步：多重 MD5 指纹消重

①从数据库供求信息表中读取供求数据需处理的各个字段值并连接生成字符串；

②对 1 中生成的字符串进行分词；

③对分词结果统计词频，按照词频高低重新生成一个新的字符串数组；

④对字符串数组进行 MD5 编码，得到 MD5 编码数组，写入数据库中供求信息临时表；

⑤对数据库中供求信息临时表建立索引；

⑥按照时间段在索引文件中进行 RangeQuery 查询产生结果集 hits1；

⑦取出 hits1(i)的第一条记录 hits1(0)，按照 MD5 编码数组中前三个元素值为关键字在索引文件中进行 TermQuery 查询生成结果集 hist2；

⑧如果 hist2.length()>1，说明可能有相似记录，计算两条数据中含有的相同 MD5 指纹数 s；

⑨如果 s>t(t 为设定的阀值)，则认为两条数据记录相似，得到相似记录的 ID，在数据库中删除，并在索引文件中同步删除；

⑩i++，返回步骤 7 中取下一条记录，直到 hits1(i)记录为空结束。

3 实验结果及分析

本文从原始数据库中（网络爬虫采集到的，但是没有经过任何处理的数据）抽取 70 万条农业供求信息数据，利用本文介绍的方法进行试验，然后分别使用 2.2 算法描述中的步骤一和步骤二进行独立实验，并比较这三种实验结果的查准率和召回率（也称查全率）以及响应时间。

3.1 查准率

查准率反映了的算法所发现的近似数据中有多少是正确的近似数据结果，假设算法检测到了 S 个重复数据记录，其中个正确结果，即符合我们对于近似数

据的定义。则算法的查准率为 Precision = p =S_o/S，计算查准率本文使用的方法为：在大数据的实验结果中进行采样，进行人工评测，取样本的准确率的平均值作为算法的查准率[4]。

3.2 召回率

召回率是算法所发现的正确的近似数据记录占全部近似数据记录的百分比。计算召回率方法为：由于不能确定实际全部近似数据记录的个数，用算法发现的近似数据记录集合的大小与实验中所有算法得到的近似数据记录的并集的大小之比来代替查全率。假设算法检测到了个近似数据记录，而数据集中实际存在 S 个近似数据记录，则算法的召回率为：Recall = r = S_o/S_o

3.3 结果

运行环境：PC 机，2.4GHZ CPU4，2G RAM，Windows XP 操作系统，JAVA 实现算法。在我们的垂直搜索引擎“搜农”上实验得到结果如表一所示：

表 1　查准率、召回率及响应时间

	查准率	召回率	响应时间（ms）
方法一	95.1%	96.3%	50822974
方法二	98.7%	89.5%	34994843
方法三	90.2%	95.8%	131031219

方法一为使用本文提出的先使用单 MD5 指纹进行完全相同数据记录消重，然后对消重后的数据库利用多重 MD5 指纹进行相似数据消重；方法二为单独使用单 MD5 指纹进行消重；方法三为单独使用多重 MD5 指纹进行消重。为了方便对比作出查准率和召回率对比图，如图 1 所示。

实验结果表明：在查准率和响应时间上，方法二最高；在召回率上，方法一最优；考虑到方法二不能有效消除近似的数据记录，而只能消除完全相同的数据记录，所以本文选用方法一。

4 总结

本文介绍了如何利用 MD5 指纹技术对大型数据库中的重复以及相似数据进行清洗，并通过实验得到查准率，召回率和响应时间。结果表明本文所提出的

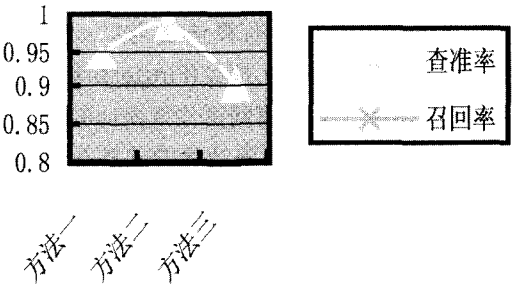

图 1　查准率与召回率对比图

方法能很好的满足实际要求。进一步的工作是：如何消除同义词构成的相似数据。比如在农业词汇中，西红柿，番茄以及杨柿子是代表统一品种，在农业价格数据中由于文档内容较少，这时品种名在数据相似性程度上就起到了至关重要的作用。

参考文献

1 Brooder A,GlassmanS,ManasseM, ZweigG. Syntactic clustering of the Web.// Proceedingsof the 6th Int'l World Wide Web Conf.(WWW) ,1997:391-404.

2 Cho J,Shivakumar N,Garcia-Molina H "Finding replicated Web collections." In Proceedings of 2000 ACM International Conference on Management of Data (SIGMOD), May 2000.

3 Rivest RL.The MD5 Message-Digest Algorithm, Request for Comments1992:1321,April.

4 王建勇,谢正茂,雷鸣,李晓明.近似镜像网页检测算法的研究与评测,电子学报,2008,28(11):130-132.

5 Shivakumar N,Garcia-Molina H,SCAM: A Copy Detection Mechanism for Digital Documents. // Proceedings of the 2nd International Conference on the Theory and Practice of Digital Libraries(DL'95), 1995.