第 35 卷 第 3 期 计 算 机 工 程 2009 年 2 月
Vol.35 No.3 Computer Engineering February 2009
·软件技术与数据库· 文章编号：1000—3428(2009)03—0023—03 文献标识码：A 中图分类号：TP391

结构化信息的去重方法

李 林，刘桂峰，赵朋朋，崔志明
(苏州大学智能信息处理及应用研究所，苏州 215006)

摘 要：针对载有结构化信息的网页，提出一种基于学习的去重方法。通过先期准备的样本定义分类器，根据分类器对页面中结构化信息的不同属性字段进行分类和距离计算，计算出整个信息对象和已分类样本信息的距离，以这些距离与阈值的大小关系判断该信息对象是否重复。

关键词：相似性测度；去重；聚类

Duplication Deletion Method for Structural Information

LI Lin, LIU Gui-feng, ZHAO Peng-peng, CUI Zhi-ming
(Institute of Intelligent Information Processing and Application, Soochow University, Suzhou 215006)

【Abstract】This paper proposes a learning-based duplication deletion method for structural information on Web. It prepares a training set for producing classifier, classifies different attribute fields of structured information in pages, and computes the distances according to the classifier. The distance between the whole information object and classified sample information is computed, and whether the record is duplicate by comparing with threshold is judged.

【Key words】similarity measure; duplication deletion; clustering

1 概述

近年来，网络技术和规模得到很大的发展，网络中的信息量空前的巨大，其扩大速度也是前所未有，因此，当前互联网的信息处理中最大的问题是页面信息的大量冗余。其解决方法就是去重。

观察页面的形式可以发现，网页中的信息分为 2 种：载有结构化信息的网页和载有非结构化信息的页面。载有非结构化信息网页的冗余、重复主要是由用户对网页的转载造成的；载有结构化信息的网页的重复在一些商业销售性网站上出现较多。在网络上检索信息时，网页信息的大量重复往往导致检索效率极其低下。

在非结构化页面去重方面已经出现了一些方法，但在载有结构化信息页面的去重方面，方法还很少。这方面的去重研究具有很大的实际应用价值，因为在实际生活中，结构化信息的应用范围非常广，而且对信息进行多维度描述时，一般结构化描述将更简洁、准确，所以有必要对结构化信息进行整理、收集。互联网所含信息量极大，其信息也更全面，可以从中检索出大量有用的结构化信息。但由于网络中存在大量重复的载有结构化信息的网页，因此需要对其进行去重。如商品房信息的发布，如果多个网站上有同一个房产的销售信息，那么在对多个网站进行检索的过程中，就将出现大量相同的房产信息。

本文提出了一种基于学习的网页结构化信息去重方法，与现有的一些方法相比，具有一定的补充作用。

2 相关的研究工作

目前大部分网页去重方法集中在对非结构化页面的去重，主要分为 2 大类：基于 URL 的去重和基于网页文本内容的去重。

基于 URL 的去重方法主要基于哈希函数对网页 URL 进行管理，以判断 2 个 URL 是否相同[1]，这种方法的执行速度很快，对于那些对准确率要求不高但要求快速的去重任务具有很好的效果。相比之下，基于网页文本内容的去重在方式选取上比较多样。在这类方法中，基于特征串和基于词频聚类 2 种方式较为常用。基于网页特征串的方法[2-3]，主要是按照一定规则，先在网页的不同位置选取特征码，按顺序组合成特征串，作为网页的标识，在去重的过程中，通过比较网页的特征串达到网页去重的目的。基于词频聚类的方法[4]先对网页中的文本进行分词，再计算每个词在文本中出现的词频，然后根据设定的规则比较网页中相同词的词频，进行去重。还有一些基于内容的方法，如通过计算页面内容重合度判断它们的相似关系[5]。但以上方法都无法处理载有结构化信息的页面。本文提出的去重方法可以弥补这方面的不足，使得去重的网页类型范围更广。

3 去重方法的设计

本文的去重方法有 2 个必要步骤：(1)在去重开始时，对不同网站结构化信息模式的字段之间进行映射。不同网站之间信息的结构模式往往有一定差距，所以，必须先处理不同网站间信息的字段映射问题。(2)在字段映射后对具体的记录进行计算分值，然后根据记录的分值进行去重，包括分类器

基金项目：国家自然科学基金资助项目(60673092)；2005 年度教育部科研基金资助重点项目(205059)；2006 年江苏省“六大人才高峰”基金资助项目(06-E-037)；2006 年度江苏省软件和集成电路业专项基金资助项目([2006]221-41)；2007 年度江苏省研究生创新计划基金资助项目(CX07B-122z)

作者简介：李 林(1982－)，男，硕士研究生，主研方向：数据挖掘；刘桂峰，硕士研究生；赵朋朋，博士研究生；崔志明，教授、博士生导师

收稿日期：2008-07-20 **E-mail：**lilin1030id@yahoo.com.cn

的生成和记录的去重 2 步。

3.1 分类器的生成

本方法包括 2 个分类器：字段映射分类器和字段值分类器。分类器是通过训练预先提供的样本集产生的，训练样本集包括从多个网站抽取出的不同样本信息，允许其中出现重复的样本，因为在字段值分类器的生成过程中，可以根据重复样本数量为字段赋权值。

3.1.1 字段映射分类器的生成

对字段进行映射是进行结构化信息去重的预备步骤，其映射过程是通过字段映射分类器完成的，分类器的目的是对信息的属性字段进行分类、归类。

首先通过无监督的分类方式对样本集进行聚类，建立不同模式字段之间的映射关系。因为每个网站中描述信息的字段数目、划分都不一样，所以开始必须建立不同网站信息模式之间的字段映射。这种映射关系的确定可以应用于每个网站的数据集，通过字符特征判断、长度的比较、同义词的统计[6]等方式实现。

设 2 个网站数据集的信息模式为 $X(Dx_1,Dx_2,\cdots,Dx_n)$和 $Y(Dy_1,Dy_2,\cdots,Dy_m)$，其中，Dx_i 和 Dy_i 分别是 X 和 Y 的属性字段，它们对应的具体记录向量分别设为 $\boldsymbol{x}(dx_1,dx_2,\cdots,dx_n)$和 $\boldsymbol{y}(dy_1,dy_2,\cdots,dy_m)$。将网站信息模式中的字段按字段的字符特征进行分组，比如，商品编号和出版日期的字符特征为数字表现形式，而商品名称和出版社大多是中文或英文的拼写形式。分组是一个不可或缺的步骤。其目的是只在具有相同字符特征的分组之间进行字段比较，从而提高比较效率。之后计算两两具有相同字符特征但属于不同网站模式的字段之间的距离，进行相似性测度，其中，字段 Dx_{pi} 拥有 m 个字段值；Dy_{qj} 有 n 个。将每个字段看成多维向量，维数等于字段值个数，那么两两字段间的距离应用欧几里德距离进行计算：

$$dist(Dx_{pi},Dy_{qj})=\sqrt{\sum_{k=1}^{m}\sum_{l=1}^{n}((Dx_{pi})_k-(Dy_{qj})_l)^2} \tag{1}$$

其中，k 和 l 分别表示 2 个字段下第 k 个和第 l 个字段值；p 和 q 分别表示 2 个字段所在的字符特征分组。再设定一个阈值 DS ，将这些距离中值最小且小于 DS 的距离对应的两个字段归入一类。在剩下的字段中，将距离最小的归入一类，依此循环，直到其中一个分组中没有字段或没有比 DS 小的距离为止。该聚类过程可以扩充到所有分组，使所有模式的字段都被分类到一个类中(类中不同的字段属于不同的网站)，每个字段都与一个类映射。如果所有的网站两两进行字段映射后还有字段未被聚类，那么该字段就是一个极端的孤立点。

依上面的步骤，生成字段映射分类器的工作完成。

3.1.2 字段值分类器的生成

字段值分类器的生成是承接在字段映射分类器之后的。当样本中的结构化信息建立好字段到类的映射后，就对字段对应的类赋以权值，权值体现了模式中不同的字段在分类信息时的作用。结合字段的权值，采用聚类的方法计算数据集中每条信息的分值，然后根据这些分值计算记录间的距离来判断是否重复。字段值分类器的生成由以下 5 个步骤实现。

(1)为字段对应的类确定权值。在进行记录聚类时，记录间的距离是根据记录的所有字段值计算的。在所有字段中，往往只有一部分字段对分类具有很大的影响，其他字段的影响则较小，如果不加区分地依赖所有字段进行相似性度量，则会使记录间距离的计算被许多不相关的字段所支配，如果忽略这些，就可能出现实际不相似的记录被判定为相似的情况，在很大程度上导致维度灾难的出现。为了避免这种情况，为每个字段赋以权值。因为不同网站信息对应的字段都映射到不同的类中，所以可以直接对类赋权值，字段的权值就是所映射的类的权值。

权值根据 2 个因素确定：该类下所有字段值中出现词频最大的词频值，类中字段的个数。如果类的最大词频比较大，说明这个类下重复的字段值太多，对计算记录间距离的影响较小，那么将该类的权值设置得较小；反之，为那些最大词频较小的类赋以更高的权值。类中字段的个数也影响权值的确定。类中的字段个数少是因为有的网站没有在信息中设置该类字段，这也说明该字段不是必需的，其重要性不大，权值应该设置得较小；反之，重要性就大，其权值也要设置得较大。每个类的权值 C_j 可以通过下式计算：

$$C_j=(\frac{1}{m+1}\times j+W_j)\times Hr,\ 1\leqslant j\leqslant m \tag{2}$$

其中，j 为该类在所有类中的序号，Hr, m 为常数；m 为类的个数；Hr 为要进行去重的数据集中记录的个数；W_j 体现当前类的重要性，并将重要性传递给 C_j 的量，其取值范围是 1~10。

在式(2)中，代入量 W_j 的计算过程如下：

先定义一个量 $Rank_j$：

$$Rank_j=\frac{Cnum}{fc}+Cnum\times\frac{Maxc}{avg(\sum_{i=1}^{n}fc_i)} \tag{3}$$

其中，fc 表示当前类下相同字段值的最大词频；$Cnum$ 表示当前类中字段个数；$Maxc$ 表示拥有最大字段数的类中的字段个数；n 为拥有最大字段数的类的个数；fc_i 为拥有最大字段数类的最大词频，$avg(\sum_{i=1}^{n}fc_i)$ 是所有 fc_i 的平均值。

W_j 的计算如下：

$$W_j=10(\frac{Rank_j-\min(Rank_i\mid i=1,2,\cdots,m)}{\max(Rank_i\mid i=1,2,\cdots,m)-\min(Rank_i\mid i=1,2,\cdots,m)}) \tag{4}$$

可以看出，W_j 的取值在 1~10 之间，为的是让权值 C_j 体现类的重要性，同时减小 C_j 的数量级，从而减小运算的难度。

由于 W_j 受取值范围的限制，将出现许多类的 W_j 是相同的情况，因此将 $1/(m+1)\times j$ 代入式(2)将完全消除类权值相同的现象。显然，$1/(m+1)\times j$ <1，且不同的 W_j 之间的差是整数，这样可以使权值 C_j 体现类的重要性，而各个 C_j 的值之间有明显的差别。代入 Hr 则是为了突出 C_j 对式(6)的影响。W_j 的重要性是由 Fv 传递的，而在计算 Fv 的式(3)中，影响类重要性的 2 个因素分别是 fc 和 $Cnum$。因为不同网站的信息模式结构不同，所以在某些类中映射的字段就较少，该类下重复字段值的数量也大大减少。为了减小这类情况的极端孤立点对计算 $Rank_j$ 的影响，使用 $Cnum/fc$ 来平衡这些孤立点的影响。代入 $Maxc/avg(\sum_{i=1}^{n}fc_i)$，目的是使 $Cnum$ 扩大到和 fc 相近的数量级上，$Maxc$, $avg(\sum_{i=1}^{n}fc_i)$ 也是为了减小孤立点的影响。

(2)应用聚类方法对所有字段值进行分簇。通过检索字段值计算同一类下字段值之间的距离，将其作为进行分簇的依据。对字段值的聚类处理是以类为单位进行的，同时设定一个阈值 Fv，用来判断是否把字段值归入当前簇。

在聚类的开始，把第 1 个字段值保存为一个簇的中心。在检索下面的字段值时，要计算当前字段值 R_j(j 为字段值的下标)和所有已存在的簇中心的距离 D_{ij}(i 类的索引)。如果 D_{ij}

大于阈值 Fv，则把 R_j 作为一个新簇的中心；如果 D_{ij} 小于 Fv，则把 R_j 归入计算后距离最小的簇中心所在的簇。分簇过程如图 1 所示。

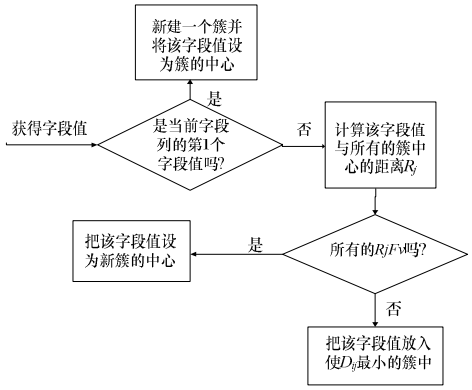

图 1　字段值分簇过程

Fv 的设定影响到同一字段下分簇的精细程度，设置得过小，一些相似的字段值就可能被划分到不同的簇中，导致某些重复记录未被识别。同时分簇过多会使字段值与簇中心的比较次数增多，从而耗费很多资源。如果设置得太大，就会把许多不重复、不相似的字段值划分到相同的簇中，导致去重时去掉很多不重复的记录。

(3)计算所有簇中不同字段值在簇中对应的分值——簇分值，记为 M_{ij}。M_{ij} 通过下式获得：

$$M_{ij}=k\times Hs+D_{ij} \tag{5}$$

其中，Hs 也是一个常数；k 为簇的序号下标。在计算簇分值时，k 可以起到区分不同簇的作用。为了强化 k 对计算结果的影响，设 $SMax$ 为当前类下元素个数最大的簇的元素个数，Hs 是以 $SMax$ 为自变量的线形函数值，再代入式(5)，就可以很好地区分同一类下不同簇的分值。

(4)获得所有单个字段值的簇分值 M_{ij} 后，该字段值的完整分值为：C_j+M_{ij}。那么一个记录的总分值就是它所有字段值完整分值的和：

$$Record_i=\sum_{j=0}^{n}(C_j+M_{ij}) \tag{6}$$

其中，n 为记录中字段的个数；i 为记录的下标；j 为字段的下标。

(5)判断重复记录。获得各个记录的分值 $Record_i$ 后，就可以通过对比记录的分值计算它们之间的差值，判断是否是重复记录。可以预设一个差额阈值 Fr，将 2 个记录分值差值的绝对值和这个差额阈值进行比较，只要满足 $|Record_i-Record_j|\leqslant Fr$，就认为这 2 个记录是重复的，可以删去其中一个。Fr 可以根据具体情况设置，如果精度要求很高，可以将差额阈值设为 0，如果精度要求不太高，或要求排除后获得的数据有明显区别，则可以将这个差额阈值设得较大。

经过上面的步骤，就完成了字段值分类器的生成。

3.2　结构化信息的去重处理

结构化信息的去重处理也需要 2 个步骤：字段映射，记录的分值计算。

生成分类器后就可以用分类器依照本文的方法进行字段映射。将当前获得的网站模式中的字段按字符特征进行分组，计算每个字段和分类器中字段类的距离，然后将字段归入距离最小且小于阈值 DS 的类中，其过程与上面所述大致相同。

计算记录的分值也是在建立字段映射之后完成的，其过程类似于字段值分类器的实现过程。先计算每个字段的权值 C_j，接着根据已有的簇中心计算各个字段值对应的簇分值 M_{ij}。与分类器生成过程相似，对于那些与所有的簇中心距离大于阈值的字段，将其作为新的簇中心。在计算每个记录所对应的 $Record_i$ 后，通过比较 2 个记录的 $Record_i$ 是否满足 $|Record_i-Record_j|\leqslant Fr$，以判断 2 个记录是否是重复记录。

4　实验及结果分析

4.1　实验设置

对上述方法进行检测，实验前先准备一个图书信息的数据集(数据集中的图书信息是从多个图书网站爬取获得的，含有大量重复的记录)，其中包括 60 多万本书的记录。根据实验的需要，从中选择部分字段，包括书名、作者、出版社、价格、出版时间、版次、内容提要。

4.2　实验结果

依照本文方法计算所得每个字段的权值如表 1 所示。

表 1　每个字段分值表

	书名	作者	出版社	价格	出版时间	版次	内容提要
j	1	2	3	4	5	6	7
$Record_j$	7	5	4	4	4	2	1
C_j	7.125	5.25	4.375	4.5	4.625	2.75	1.875

之后对每个字段下所有的字段值进行分簇。

将 2 个阈值设为：Fv=2, Fr=2，计算结果如表 2 所示。

表 2　分簇统计表 1

	书名	作者	出版社	价格	出版时间	版次	内容提要
簇数	308 565	598 487	523	1 878	612	6	599 476
最大簇的元素个数	9	12	7 381	2 449	423	450 306	3

统计得到重复记录有 306 786 条。

将 2 个阈值设为：Fv=1, Fr=2，计算结果如表 3 所示。

表 3　分簇统计表 2

	书名	作者	出版社	价格	出版时间	版次	内容提要
簇数	316 691	598 493	523	1 878	612	6	599 481
最大簇的元素个数	7	12	7 381	2 449	423	450 306	3

统计得到重复记录有 298 676 条。

将 2 个阈值设为：Fv=2, Fr=0，计算结果如表 4 所示。

表 4　分簇统计表 3

	书名	作者	出版社	价格	出版时间	版次	内容提要
簇数	316 691	598 493	523	1 878	612	6	599 481
最大簇的元素个数	7	12	7 381	2 449	423	450 306	3

统计得到重复记录有 326 113 条。

4.3　实验分析

从上面的统计信息可以看出，当 Fv 变小时，重复记录数减少，反之增多；当 Fr 减少时，重复记录数增多，反之减少。

实验目的是考察去重的正确性，因为人为地判定去重的正确性很难，所以针对以上 3 组实验分别随机抽取了 20 组网页，每组 50 个网页，然后进行人工判别。结果证明，错分的网页比例不大，去重的准确率达到 89.3%，招回率达到 88.9%。错分的主要原因是网页中结构化信息字段的复杂性造成了字段分类时出现错误。

5　结束语

页面中结构化信息的去重具有很大的实际应用价值和空间。本文的方法具有很大的实用性。它通过分类和计算并对

（下转第 28 页）