第35卷 第20期 计 算 机 工 程 2009年10月
Vol.35 No.20 Computer Engineering October 2009
·开发研究与设计技术· 文章编号：1000—3428(2009)20—0276—02 文献标识码：A 中图分类号：TP391

面向交通管理领域的分类索引算法

李云鹏，熊桂喜
(北京航空航天大学计算机学院，北京 100083)

摘 要：通过计算数据集与样本集在主题特征向量上的相似度对数据集进行信息筛选与分类处理，以便有效地组织和分析交通管理领域内的数据资源，使查询结果分布在最相关的数据集中。利用 Hadoop 分布式应用程序框架使各工作节点协同完成索引的构建。根据用户查询的类别，只在最相关主题索引库中进行查找，以提高检索效率。

关键词：主题特征向量；分类索引；交通管理

Traffic Management Domain-oriented Classified Index Algorithm

LI Yun-peng, XIONG Gui-xi
(College of Computer, Beijing University of Aeronautics and Astronautics, Beijing 100083)

【Abstract】 To effectively organize and analyze data within traffic management, this paper makes use of computing similarity between data set and sample set topic feature vector to filter and decide which categories the data sets belong to. All work stations work together to build index with the help of Hadoop distributed application framework. According to the category of query, the algorithm only consults correlative topic index for results to achieve a better retrieval performance.

【Key words】 topic feature vector; classified index; traffic management

1 概述

交通管理领域内的综合信息管理和一般的搜索引擎有着类似的结构，是针对交通这一特定领域需求，分区域、按主题创建索引库的专业搜索引擎。在交通管理的日常工作中，所涉及的基础数据覆盖了车辆管理、交通设施、交通监测、指挥调度、信息发布等多类业务系统，这些数据来源分散，且处于相互独立的状态，给信息的共享和检索带来不便。从组织方式出发，数据可分为关系数据库中的结构化数据、网站所发布的半结构化数据、日常应用中所产生的文件与电子表格等数据，呈现多样化的特点。因此，对这些宏大分散的异构数据源进行管理是实现信息高度集成与共享的前提。

目前，面向主题的索引构建技术主要分为 2 种：(1)基于内容的方式，通过建立针对主题的词表，对搜集的信息进行索引，词表的建设也越来越多引入知识表示的方法。(2)基于链接分析的方式，通过对链接进行分析，找出各个数据集之间的引用关系，按引用关系对信息进行分类后构建索引[1-2]。

本文在基于内容的主题索引技术基础上，对所采集的数据集进行筛选、标注关键词、建立主题索引库，提供统一的信息检索接口，将车辆管理、交通监测、交通执法、法制宣传、事故处理等业务关联起来，帮助用户快速准确地定位所需的信息，实现闭环执法和闭环交通管理机制。

2 分类索引算法

分类索引算法按各自的功能分为构建主题结果集和建立索引库 2 个部分。对于给定的不同类别专业样本集，在经过中文切分词和特征词提取后，以向量的形式表示，作为该类别的主题特征向量。在确定类别的过程中，一个待分类的数据集同样在经过中文切分并表示成向量后，应用分类算法计算出该数据集与不同类别的主题特征向量之间的相似度，再与各个类别的阈值作比较，类别的阈值由自定义的阈值策略预先确定。选择相似度大于阈值的类别，作为该数据集的分类结果。

经过数据资源筛选与分类后的主题结果集由 Hadoop 分布式文件系统进行统一管理，利用 Hadoop 的分布式应用程序框架，将分布式文件系统内的数据集合计算分布在各个工作节点上，各节点协同工作完成资源的分类索引。在索引库构建过程中，将关键词的标注信息添加到不同主题的索引库中，为检索时的查询关联提供底层支持。

3 主题结果集的构造

3.1 关键词权重的计算方法

由于关键词向量具有灵活、用户习惯使用和不易漏检信息的特点，因此使用加权关键词向量代表数据集的模式[3]。根据交通领域的专业样本集 D(如交通拥堵)，提取集合 D 中关键词总数 t，设 $K=\{k_1,k_2,\cdots,k_n\}$(如事故、施工、快速路、追尾、并线、出入口、拥堵、高峰)是 D 中所有关键词的集合，任一校验集 D_j 可以表示成 t 维空间 R^t 中的一个向量：$\boldsymbol{Vector}(D_j)=\{W_{1j},W_{2j},\cdots,W_{tj}\}\in R^t$，其中，$W_{ij}\geq 0$，表示关键词 K_i 在 D_j 中的权重。

计算数据集 D_j 中每个关键词的权重，传统的关键词得分加权方法运用 TF-IDF 公式 $W(t,d)=f(t,d)\times \mathrm{lb}(N/n_t)$，其中，$W(t,d)$ 为关键词 t 在文档 d 中的权重，而 $f(t,d)$ 为关键词 t 在文档 d 中的词频，即词 t 在文档 d 中出现的总次数，N 为专业样本集的文档总数，n_t 为样本集中出现关键词 t 的文档数[4]。另外，数据集合中的每个关键词权值大小还与关键词

基金项目：“十五”国家科技攻关计划基金资助项目“现代中心城市交通运输与管理关键技术研究”(2005BA414B04)

作者简介：李云鹏(1985－)，男，硕士研究生，主研方向：信息检索；熊桂喜，副教授

收稿日期：2009-03-27 **E-mail：**yunpeng.flying@yahoo.com.cn

的长度和出现在页面中的位置有关。例如，一个关键词出现在标题比出现在正文中更能代表该数据集的内容，那么可以认为它的重要性高于正文中出现的关键词。在综合考虑关键词的词长与关键词在页面中的位置信息后，每个关键词新的权重计算公式如下：

$$W_{ij} = W(t,d) \times Len\_Weight \times Pos\_Weight$$

其中，Len_Weight =lb (关键词的长度)；Pos_Weight 为位置的权值(设置的参数)。

3.2 主题特征向量的计算方法

根据聚类分析的基本原理，可以由样本集中的平均关键词权重来刻画一个类别的总体特性，称为主题特征向量。围绕人(管理者与被管理者)、车(机动车与非机动车)、路(快速路、主干路和支路)诸方面因素，交通信息发布的类别主要分为交通流信息、交通拥堵信息、交通法规信息、违章车辆信息[5]。

设该类别的样本集 $D=\{D_1,D_2,\cdots,D_n\}$，主题特征向量 $\boldsymbol{Vector}(C) \in R^t$ 可按如下公式计算：

$$\boldsymbol{Vector}(C)=\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{Vector}(D_j)=(\frac{1}{n}\sum_{j=1}^{n}W_{1j},\frac{1}{n}\sum_{j=1}^{n}W_{2j},\cdots,\frac{1}{n}\sum_{j=1}^{n}W_{tj})$$

3.3 分类算法

依次计算数据集 D_j 与各个类别的主题特征向量的相似度。相似度根据 2 个向量之间的内积来计算，对于任意的 $D_j \in D$，其相似度计算方法如下：

$$Sim(D_j,C)=\frac{\sum_{i=1}^{t}(W_{ij}\times\frac{1}{n}\sum_{m=1}^{n}W_{im})}{\sqrt{\sum_{i=1}^{t}W_{ij}^2}\sqrt{\sum_{i=1}^{t}(\frac{1}{n}\sum_{m=1}^{n}w_{ij})^2}}$$

根据试验数据，阈值可选择 60%，即认为相似性距离大于 60%的文档是主题相关的。在一个数据集被判断属于交通领域资源后，再比较其与每个类别主题特征向量之间的相似度，将该数据集分配到相似度最大的类别中。

4 主题索引库的建立

在对信息资源进行筛选、分类后，需要对这些不同主题的结果库建立索引，使系统能够根据用户查询的类别到对应主题的索引库中进行检索，从而加快查询速度。本文借助 Hadoop 分布式应用程序框架实现对大规模数据可靠有效的并行处理。该框架提供了一组稳定、可靠的接口，其关键在于配置 Map/Reduce 任务。

4.1 Map/Reduce 编程模型

Map/Reduce 是一种用于处理大规模数据的编程范式，是函数式编程上的概念。Map 将过程应用于数据以产生新的数据，Reduce 则是将数据进行归并。这种模型把大量复杂的分布式计算看作是一系列建立在 Key/Value 映射数据集上的操作[6-7]。

Map/Reduce 编程模型把对数据集的所有操作都归结为 2 个阶段 Map 和 Reduce。首先，输入数据被分割成有序的数据块，工作节点应用自定义 Map 函数到每个数据块，得到中间数据集。然后，中间数据集被送往 Reduce 阶段的工作节点，应用自定义 Reduce 函数对其进行去重、过滤等后期处理，最后得到需要的结果。其过程原型如下：

Map　　(k1,v1) → list(k2,v2)

Reduce　　(k2,list(v2)) → list(v2)

4.2 索引库的生成

本文运用 Hadoop 作为分布式系统的框架，以 Lucene 作为建立索引的工具，构建主题索引库，其构建的总体过程见图 1。

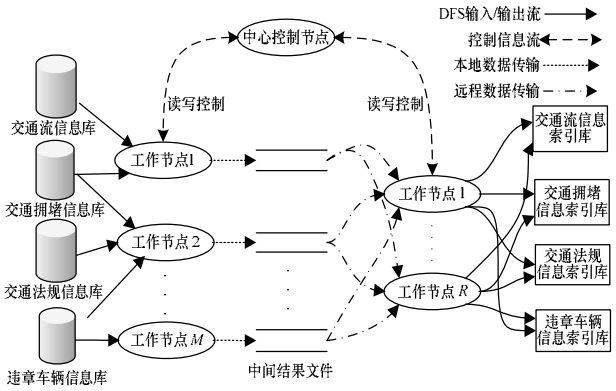

图 1　不同主题的索引库的构建过程

(1)数据准备阶段

不同类别信息库中的数据以一定的格式存放数据，系统框架依赖于所定义的输入数据格式一方面验证待索引文件的数据格式是否正确；另一方面将待索引文件集分割成以文件为单位的多个数据块，然后将这些划分好的数据块分配给各个工作节点。同时，提供读取数据项的接口 Key/Value 序列对作为数据块信息，作为各个任务的输入数据。

(2)关键词的标注

数据存储在某一工作节点上，就由该节点进行此部分数据的计算，这样可以减少数据在网络上的传输，降低对网络带宽的需求。定义页面元数据类，从本地数据块中获取页面文档在 Hadoop 分布式文件系统中的位置，读取页面信息，加入关键词的标注信息，保存至元数据类，生成新的 Key/Value 序列对。同时将中间结果数据的位置信息告诉中心控制节点，为接下来的步骤做准备。Key 值由资源 ID 经过 MD5 算法处理后生成的字符串形式表示；value 值由解析提取关联后的文档文件生成的元数据类表示。

(3)类别信息加入索引文件

工作节点先从中心控制节点得到中间结果数据在 Hadoop 分布式文件系统上的位置信息，并用远程过程调用从文件系统上的局部磁盘中读取数据。在读取所有数据后，工作节点按照 Key 值进行排序，遍历排好序的中间数据，对具有相同资源 ID 的记录集筛选出得分最高的记录，作为下一阶段的输入数据，被解析后分成各个不同的域存入索引文件。

分词后需要进行索引的域信息如文档摘要、类别信息、内容信息经过索引项的抽取操作后，将划分出来的词条加入索引文件中。同样，将作为辅助信息的域，如文档链接地址、文档创建时间、关键词权重加入索引文件中。

(4)生成各类别索引库

定义输出数据格式，利用 Lucene 的读写索引类将所生成的文档信息写入索引库，将输出的各种类别的索引库写入分布式文件系统的指定输出目录中。

5 实验结果与分析

测试的样本集是实现页面自动分类的前提和基础，系统的测试样本集主要取自于北京市公安交通管理局网站上所发布的交管动态、交通法规、出行提示等官方网页信息，其分类具有一定的代表性。测试网页集中包括 139 个训练页面实例和 526 个测试页面实例。

评估页面分类的标志是映射的准确程度和映射速度。映射速度取决于映射规则的复杂程度，评估映射准确程度参照的是通过专家判断后对页面的分类结果，准确率指标是所有判断的页面中与人工分类结果吻合的网页所占的比例。

(下转第 280 页)