一种面向海量实时数据的信息检索算法

丁伟[1] 林容容 倪良胜

（东南大学计算机科学与工程系 南京 210096）

摘要：网络信息资源的迅猛膨胀推进了信息检索技术的发展和成熟，但将现有的技术应用于海量实时网络数据时，传统的信息检索算法仍存在种种不足之处。本文以CERNET 华（东）北地区的海量实时网络数据环境为依托，研究和设计了两段向量簇聚类信息检索算法，通过插入聚类和优化聚类两阶段的操作，提供高效的信息处理能力。同时，本文基于簇聚类树实现了群发邮件甄别的应用，对网络数据中的垃圾邮件进行过滤，进一步地提高检索效率。

关键字：信息检索；簇聚类；两段向量；邮件甄别

中图分类号：TP391　　文献标识码：A

随着信息产业的飞速发展，巨大的数字化信息空间使各种面向互联网信息的搜索引擎蓬勃发展，信息检索技术日益成熟。从传统的技术角度而言，搜索引擎主要分为两类：目录式分类搜索引擎（Catalog）和全文搜索引擎（Information Retrieval ，简称 IR）。前者将信息系统地加以归类，按照传统的信息分类方式来组织信息。用户可以按分类查找信息。如：全球最著名的分类搜索引擎 Yahoo。后者采用自动分词技术，提供对所搜索的页面文件中的每一个词进行查询。最典型的例子是 Digital 公司的 AltaVista。

虽然搜索引擎的核心技术已经相当成熟，但是在海量实时数据环境下，传统的搜索引擎技术仍会显得力不从心。在海量实时网络环境下，数据不停地输入到系统中，系统如何存储数据以及如何从这些海量数据中快速检索出用户所需的信息是对系统处理策略和处理能力的考验。另外，大量的垃圾电子邮件[1]易被黑客利用，占用网络带宽，将降低整个网络的运行效率。

本文针对海量实时网络数据环境下的高效信息检索展开研究，提出两段向量簇聚类算法为信息检索的查询提供高效索引，同时提出基于簇聚类的单链接算法以用于群发邮件的甄别，实现垃圾邮件的过滤[2]，提高检索效率。全文组织如下：第一节系统阐述两段向量簇聚类算法；第二节分析研究在构建簇树过程中须考虑的各属性参数；第三节介绍基于簇聚类的单链接算法；在第四节中结合实验测试结果，分析和讨论前述算法的性能；最后总结全文，展望未来研究方向。

1 两段向量簇聚类算法

1．1 基本概念定义

簇技术是一种基于统计学的技术，由聚类所生成的簇是一组数据对象的集合，这些对象与同一个簇中的对象彼此相似，与其他簇中的对象相异。簇中的数据对象可组成树状的层次结构。通过对文档的双字切词[3]处理得到词频向量[4]，以向量模型[5]来描述文档、簇的空间位置，即可利用簇树[6]组织文档。为方便讨论，首先定义如下几个概念：

（1）文档簇（Doc Cluster，简称 DC）：如果簇的元素为文档，则称为文档簇。

（2）超簇(Super Cluster，简称 SC)：如果簇的元素为文档簇或超簇，则称为超簇。文档簇和超簇统称为簇。

（3）簇树（Cluster Tree，简称 CT）：

1 丁伟，女，1963 年，教授，博士生导师，研究方向：网络行为学，网络安全，网络测量。E-mail: wding@njnet.edu.cn

通过聚类算法得到的由文档簇、超簇组成的空间层次结构称为簇树。簇树的根称为簇根（Cluster Root，简称 CR）。簇树（如图 1 所示）是一个高度平衡的树。其中，第 0 层为文档，第 1 层为文档簇 DC，从第 2 层向上为超簇。簇的高记为 H_C。

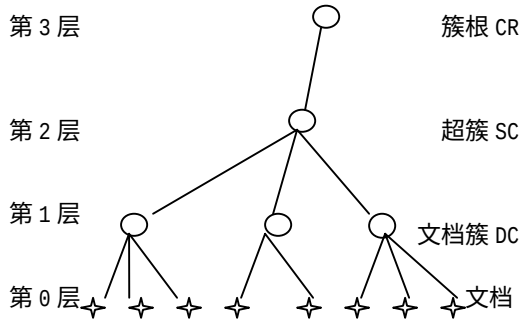

图 1 簇树的层次结构（H_C=3）

Fig 1 Structure of Cluster Tree (H_C=3)

（4）簇的质心：簇中所有元素的平均空间位置。

（5）簇的半径：该簇的聚类半径。以簇的质心为球心，以簇的聚类半径为半径所构成的球空间为该簇的聚类空间。

通过簇树来组织文档，可以将文档按照相似度进行分类，进而改善查询性能。在向量空间中，簇是一个超球体，通过簇可以将相关的文档组织在一起。

1．2 算法描述

数据挖掘中常用的两种聚类算法是划分方法和层次方法。一种综合层次聚类方法[7]——BIRCH 方法能够提供优越的聚类质量。该算法采用了一种多阶段的聚类技术，其工作分为两个阶段：

阶段一：扫描数据库，建立一个初始存放于内存的聚类特征树（CF 树）；

阶段二：采用某个聚类算法对 CF 树的叶节点进行聚类。

BIRCH 算法不仅能满足实时情况下的数据搜集效率需求，而且可以满足海量数据的聚类效果需求。但是在向量模型中，每个叶节点都是一个多维向量[8]。BIRCH 方法在处理多维向量时，仍然有一些不足之处：

- CF 树的每个节点由于大小限制只能包含有限数目，因此一个 CF 树节点并不总是对应于用户所认为的一个自然聚类。
- 随着文档向量的不断插入，从所属文档簇向上一直到 CR 的质心会不断更新，如图 2 所示，文档可能在此过程中脱离文档簇的聚类空间，产生质心漂移现象，影响聚类效果。

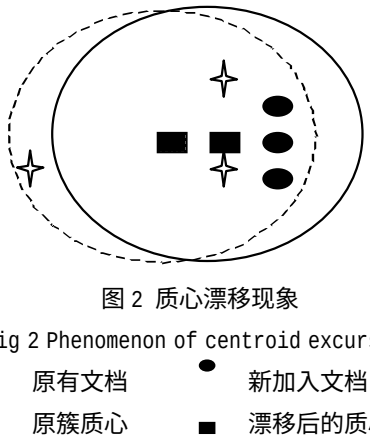

图 2 质心漂移现象

Fig 2 Phenomenon of centroid excursion

吸收 BIRCH 聚类方法的优点，结合向量模型的特点，本文提出两段向量簇聚类方法，满足海量实时数据的搜集和聚类需求。算法描述如下：

第一阶段，插入聚类——在高效收集文档的基础上实现文档的初级聚类。

在此阶段中，搜集到一篇文档后，由簇树根开始从上到下寻找和该文档最相关的叶节点（文档簇 B）。如果该文档簇 B 存在，文档就被插入到该文档簇 B 中；否则找到最相关的节点（超簇 A），从该超簇建立到最大高的一条路径，并将该文档插入到该路径的最深节点（文档簇）；如果不存在最相关的节点（即 CR 节点 A 与文档不相关），则需要重新修正 CR，并新建一条路径到最大高的一条路径，并在该路径的文档簇中插入新添加的文档。新节点插入后，由该结点所属的文档簇开始从下到上更新到 CR 的路径上的节点的质心等相关信息。图 3 为该过程的流程图。

随着文档节点的不断插入，所有文档都被插入到簇树中最相关的文档簇中，CT 的相关节点的质心向量不断更新并从下向上向传递至 CR。该过程类似 B＋树构建中的节点插入，最终形成如图 1 所示的满树。文

档被动态实时插入，聚类树被动态地构造，

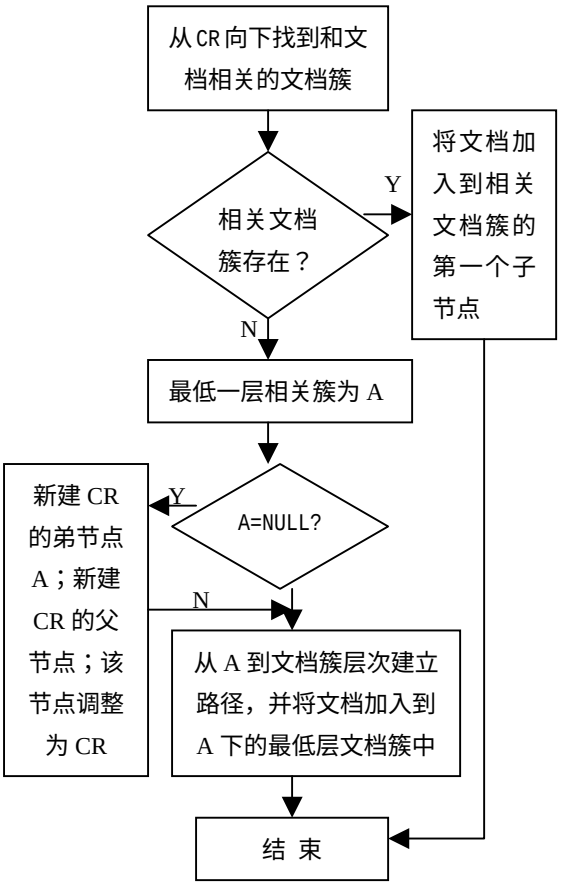

图 3 插入聚类流程图

Fig 3 Flow of Insert Clustering

所以这个方法支持增量聚类。

第二阶段，优化聚类——在遵循自然聚类的基础上消除质心漂移现象，进一步优化聚类效果。

因为质心漂移现象具有连锁效应，因此对簇树的优化聚类必须由文档簇从下到上执行：

（1）在文档簇内部逐一判断所含的文档是否漂移出聚类空间。如果是，则将该文档从文档簇中剔除，并由下到上从路径中的每一个子聚类质心中剔除文档的词频向量；然后将该文档重新插入到簇树中，并由下到上从路径中的每一个子聚类质心中增加文档的词频向量。

（2）对每一个文档簇，执行第（1）步的动作。

（3）依次对第（CH－1）层、第（CH－2）层、…、第 0 层执行第（1）、（2）步的动作。

2 簇树构建中的属性参数

在建构簇树的过程中须考虑其如下属性。

1. 节点数

簇树每个节点的空间位置由质心、聚类半径唯一决定，但是在对簇树进行查询操作时，节点数目将影响查询效率。在查询过程中，需要由簇根开始从上到下逐一比较查询词频向量和选中路径上的每一个簇的相关度，直至检索到最低层的文档簇，从而检索出与用户的查询词频向量最相关的文档簇。其查询过程大体如下：

```
S_cluster *Search(S_cluster *root, S_vector *query_vector)
{
    S_cluster *ptr;
    while(root->level != 1){
        ptr = root 所有儿子节点中与 query_vector 最相关的簇节点；
        root = ptr;
        }
    }
```

词频向量的相关度计算比较复杂耗时，为优化查询，就必须尽量减少词频向量相关度的计算次数。因而需考虑文档在簇树中的分配问题，即每个簇节点所包含的子节点数。

当簇树高 CH＝1 时，将 n 篇文档平均分配到 k 个文档簇中，词频向量相关度计算次数 $x = k + n/k \geq 2\sqrt{n}$ 。x 最小时，$k = \sqrt{n}$。由此扩展到高为 CH 的簇树：设第 k 层的节点数为 n_k，则最优情况下，

$$n_k = (n_{k-1})^{1/2} \qquad k = 1,2,3,...,(CH-1)$$

式（1）

要从簇树节点数上优化查询算法，就须使得簇树中每层节点数按照（1）式进行分配。

2. 聚类半径 rk

（1）式中对节点数分配的结论可转化为对聚类半径的确定。考虑文档向量在向量空间中平均分布的情况，设：ρ 为文档分布密度，n_k 为 k 层的节点数，簇树中文档簇聚类半径设为定值 SIM_DIST，则：

$$\because n_k = n_{k-1}^{1/2}, \qquad n_k = \rho V_k$$

$$\therefore V_k = V_k^{1/2} \qquad 又 \because V_k = \pi r_k^3$$

$$\therefore r_k = r_{k-1}^{1/2}$$

假设：r_d＝SIM_DIST，则第 k 层节点的聚类半径为：

$$r_k = r_d^{2^{k-1}} \qquad k = 1,2,3,...,(CH-1)$$

式（2）

因此，要优化簇树构建过程，就需要使簇树中所有簇节点的聚类半径按照（2）式确定。

3. 质心 c_k

文档、簇的空间位置都由词频向量来表示。根据簇树质心的定义，节点的质心向量 c_k 是该节点所包含的所有元素向量的平均值。其计算公式如下：

$$c_k = \frac{1}{n}\sum_{i=0}^{n-1} c_{ki} \qquad 式（3）$$

如果所包含的元素是文档，则 c_{ki} 为对应文档的词频向量；如果所包含的元素是簇，则 c_{ki} 为对应簇的质心向量。

4. 簇树的高 H_C

当簇树的每层节点数按公式（1）确定时，簇树的各层节点数和文档总数可通过图 4 表示。假设一个数据库所存储的文档数处于区间[65536，43 亿]中是合理的，所以按（1）式产生的簇树 H_C=5。可见只要保证簇树按最优构建，簇树的高 $H_C \leq 5$。

	文档总数	各层节点数
第 H_c-1 层	1（CR）	1
第 4 层	(1,4)	(1,4)
第 3 层	[4,16)	[4,16)
第 2 层	[16,256)	[16,256)
第 1 层	[256,65536)	[256,65536)
第 0 层	[65536,43 亿)	[65536,43 亿)

图 4 簇树各层节点数

Fig 4 The number of nodes in Cluster Tree

3 基于簇聚类的单链接算法

垃圾邮件（多为群发邮件）大大浪费了系统的处理时间和存储空间，为提高信息检索效率，本文利用信息检索技术中基于相关度计算的群发邮件甄别技术提出基于簇聚类的群发邮件甄别算法，实现垃圾邮件的过滤。

群发邮件大多数都集中在一段较小的时间内发送；另外，目前群发邮件在实时环境的含量非常高。由于单链接算法在最优情况下计算复杂度为 O（N），最差情况下计算复杂度为 O（N^2），效率较高，故可在在每个文档簇内部执行单链接算法，提高检索效率。

结合两段向量簇聚类算法，可将群发邮件甄别模块放在簇聚类的两个聚类阶段之间——完成插入聚类后，利用插入聚类的结果在文档簇内部执行群发邮件甄别，尽可能过滤掉垃圾邮件，再执行优化聚类。这样，由于过滤掉了绝大多数的群发邮件，也可以大大减轻优化聚类的压力，从而进一步增加数据处理能力。

假设共有 N 篇文档，平均分布到 K 个文档簇中，则性能提高因子[6]最小为 $\frac{N^2}{K*(N/K)^2} = K \cong \sqrt{N}$。

基于簇聚类的单链接算法利用了簇聚类的结果，不仅提高了群发邮件甄别效率，同时也利用簇聚类的存储空间代替了有效邮件池；而且由于过滤掉了群发邮件，极大地提高了簇聚类中优化聚类的效率。

4 算法性能测试

本设计的聚类过程分为两个阶段。在第一个阶段中，对于数目为 N 的数据集合，插入聚类该数据集合的计算复杂度是 O(N)，因此该算法适合实时数据的搜集。该阶段通过数据集合的单遍扫描，成了数据集合中所有数据的初级聚类，为进一步的优化聚类、群发邮件甄别奠定了基础。

在对某一小时的所有文档完成了插入聚类后，同一群发邮件组中的所有群发邮件基本被聚类在同一个文档簇中。此时在每个文档簇内部执行单链接算法，可以将同一个文档簇中的群发邮件全部甄别出来并加以过滤。因此，该小时内的群发邮件绝大多数可以被甄别出来。

优化聚类的目的主要是为了消减质心漂移现象。由于质心漂移现象具有连锁效应，因此必须从文档簇开始由下而上一直到簇根，逐层对漂移出父节点聚类空间的节点进行重新插入。优化聚类、群发邮件甄别两个模块的时间复杂度都是 O（N^2）。

为测试群发邮件的甄别性能，本设计从 CERNET 华（东）北地区网上搜集了 10GB 的网络数据。在此背景数据中混入一定量的测试数据并加以特殊标记作为群发邮件，在邮件甄别过后统计测试数据的过滤量，从而计算过滤率。

本测试中，随机选择了一份测试数据并将其数据结构 S_DATA 中的属性 fpath 赋为以“ TEST ”开头的不同字符串，向子系统发送一定数量。将该批数据时戳置为某固定小时，观察该小时聚类的过程中该组群发邮件的甄别率。为了全面考察不同长度对于群发邮件的甄别性能，选择不同长度的数据进行若干组的测试实验。

由测试结果（表 1）可知，群发邮件甄别算法的过滤率超过了 95%。

综合看来，本文设计的两段向量聚类算法可以高效实现海量网络数据环境下的实时簇聚类，基本满足了海量实时网络数据环境中的信息检索要求。

表 1 群发邮件甄别测试结果

Table 1 Result of group mail discrimination test

总数据量	54679	62882	54882	62882
输入群发邮件数	9900	9900	9900	9900
群发邮件长度	655 Byte	1364 Byte	11150 Byte	5840 Byte
含群发邮件的文档簇数	46	45	23	18
群发邮件过滤数	9854	9655	9877	9586
群发邮件漏报数	46	245	23	314
群发邮件漏报率	0.5%	2.5%	0.2%	3.17%
群发邮件甄别率	95.5%	97.5%	98.8%	96.83%

5 小结与展望

本文针对海量实时网络数据的信息检索展开研究，引入数据挖掘中的簇聚类技术，同时结合向量模型，提出两段向量簇聚类算法。该算法通过插入聚类完成对实时数据的高效收集和初级聚类，通过第二阶段的优化聚类解决多维数据聚类过程中所产生的质心漂移现象，提高对海量网络数据的实时处理能力和查询性能。同时，本文将单链接算法与簇聚类结构相结合，强化垃圾邮件的过滤，提高群发邮件的甄别效率，以节约存储空间并进一步提高信息处理效率。

本文设计的两段向量簇聚类算法虽然可以稳定高效地应用于海量实时网络环境，但仍存在诸多需改进的地方。首先，簇树聚类算法决定了簇树结构的合理性和稳定性，也决定了用户查询的效率，因此有必要在将来对聚类算法进行更深入的研究和改进。尤其是本算法中的优化聚类算法设计比较简单，在一定程度上制约了簇树结构的效率。此外，在本算法中，对于用户的查询请求，返回的是最相关的一个文档簇。但是当用户

的查询请求跨越多个簇时，简单地返回最相关的一个文档簇可能会造成查询结果的不准确。因此跨簇查询的研究也是未来的工作方向之一。

参考文献：

[1]北京晨报，垃圾邮件超过正常邮件，四部门将进行专项治理，http://www.sina.com.cn，2004-02-04.

[2]张晓东、张书杰、王万亭，“ 信息过滤的模糊聚类模型 ”，计算机工程与应用，2002，09：34 – 35

[2]XiaoDong Zhang, ShuJie Zhang, WanTing Wang. Fuzzy Clustering Model of Information Filtering. Computer Engineering and Applications. 2002,09:34-35.

[3]Ron.Papka, and James Allan. Document Classification using Multiword Features. Proceedings of the seventh international conference on Information and knowledge management. New York, NY,USA:ACM Press, 1998. 124 – 131.

[4]王斌，文本分类综述，
http://159.226.40.18/papers/
%CE%C4%B1%BE%B7%D6%C0%E0%D7
%DB%CA%F6%A3%AD%A3%AD%CD%F5%B1%F3.ppt，2002-12

[5]G.Salton, A. Wong and C.S. Yang. A Vector Space Model for Automatic Indexing. Communications of the ACM. New York, NY, USA: ACM Press, 1975, 18(11):613-620.

[6]G.Salton, M.J.McGill. The SMART and SIRE Experimental Retrieval Systems. Morgan Kaufmann Multimedia Information And Systems Series. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.,1997. 381 – 399.

[7] iawei Han, Micheline Kamber 著，范明、孟小凤译，《数据挖掘概念与技术》,机械工业出版社，2001 年 11 月

[8]颜雪松，蔡子华，一种快速聚类高维数据的算法研究，计算机工程，2003，29（1）：131

[8]XueSong Yan, ZiHua Cai, Research on a Fast Algorithm to Cluster High Dimensional Data, Computer Engineering, 2003, 29(1):131.

Information Retrieval Algorithm for Massive and Real-time Data

Ding Wei Lin Rong-rong Ni Liang-sheng

（Department of Computer Science & Engineering, Southeast University, Nanjing 210096）

Abstract: With the help of rapid expansion of information resource all over the network, Information Retrieval technology is becoming well-developed, although the current application on massive real-time data, particularly for the conventional Information Retrieval Algorithm, still have some problem. This paper investigated the massive real-time network data of CERNET, designed a Two-Phase Clustering Algorithm which is to boost the efficiency in the information management by a two-phase operation: insert clustering and optimization clustering and discussed its application in the group mail discrimination system which filtered the junk mail of network data and as a result, improved the performance of retrieval algorithm.

Keyword: Information Retrieval; Clustering; Two-phase; Mail Discrimination

附：

丁伟，女，1962 年 5 月出生，工学博士，现任东南大学计算机科学与工程系教授。研究领域包括：通用搜索引擎、PKI 证书体系和网络行为学等。