

一种面向海量实时数据的信息检索算法

丁伟 林容容 倪良胜

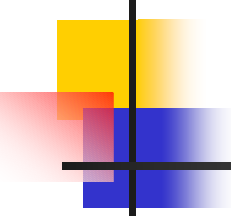
南京东南大学计算机科学与工程系



一种面向海量实时数据的信息检索算法

主要内容:

- 背景
- 两段向量簇聚类算法
- 簇树构建中的属性参数
- 基于簇聚类的单链接算法
- 算法性能测试结果



一种面向海量实时数据的信息检索算法

- 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。
- 两段向量簇聚类信息检索算法
- 基于簇聚类树实现群发邮件甄别的应用



一种面向海量实时数据的信息检索算法

利用簇树组织文档

- 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】描述文档、簇的空间位置

一种面向海量实时数据的信息检索算法

簇树相关概念定义：

- 文档簇
- 超簇
- 簇树
- 簇根
- 质心
- 半径

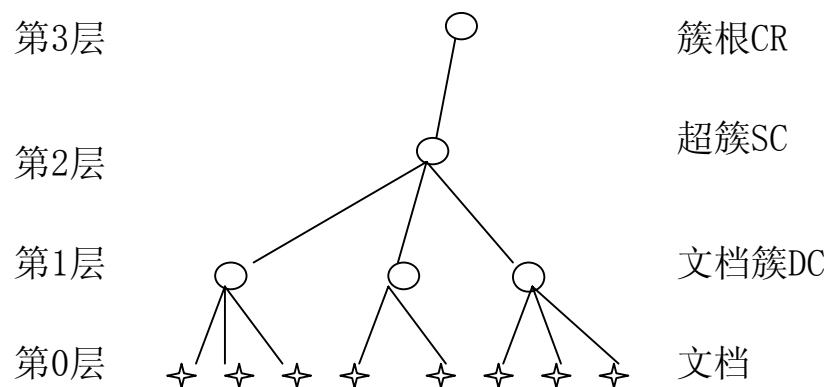


图 1 簇树的层次结构 ($H_C=3$)

一种面向海量实时数据的信息检索算法

传统的综合层次
聚类方法【7】BIRCH
方法：

- 非自然聚类
- 质心漂移

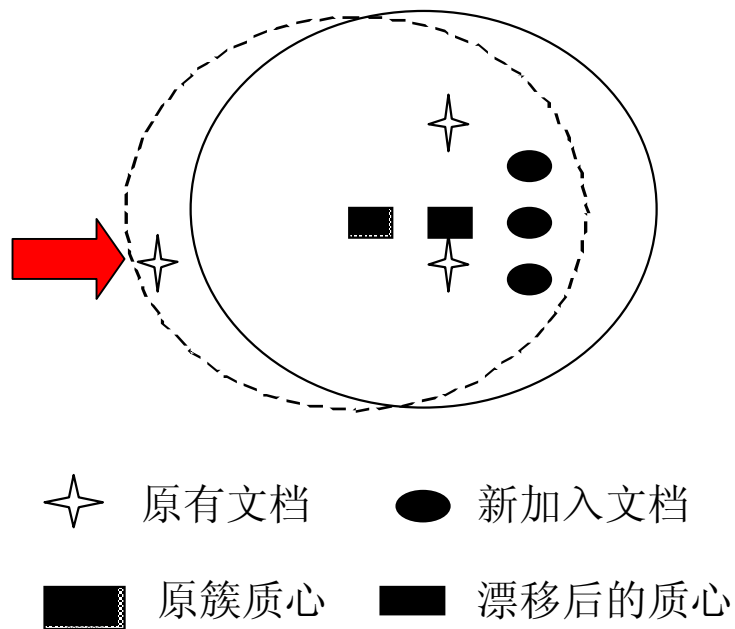
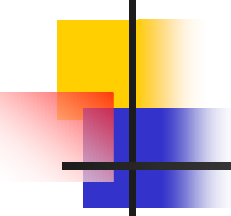


图2 质心漂移现象



一种面向海量实时数据的信息检索算法

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。

- 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。
- 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。

一种面向海量实时数据的信息检索算法

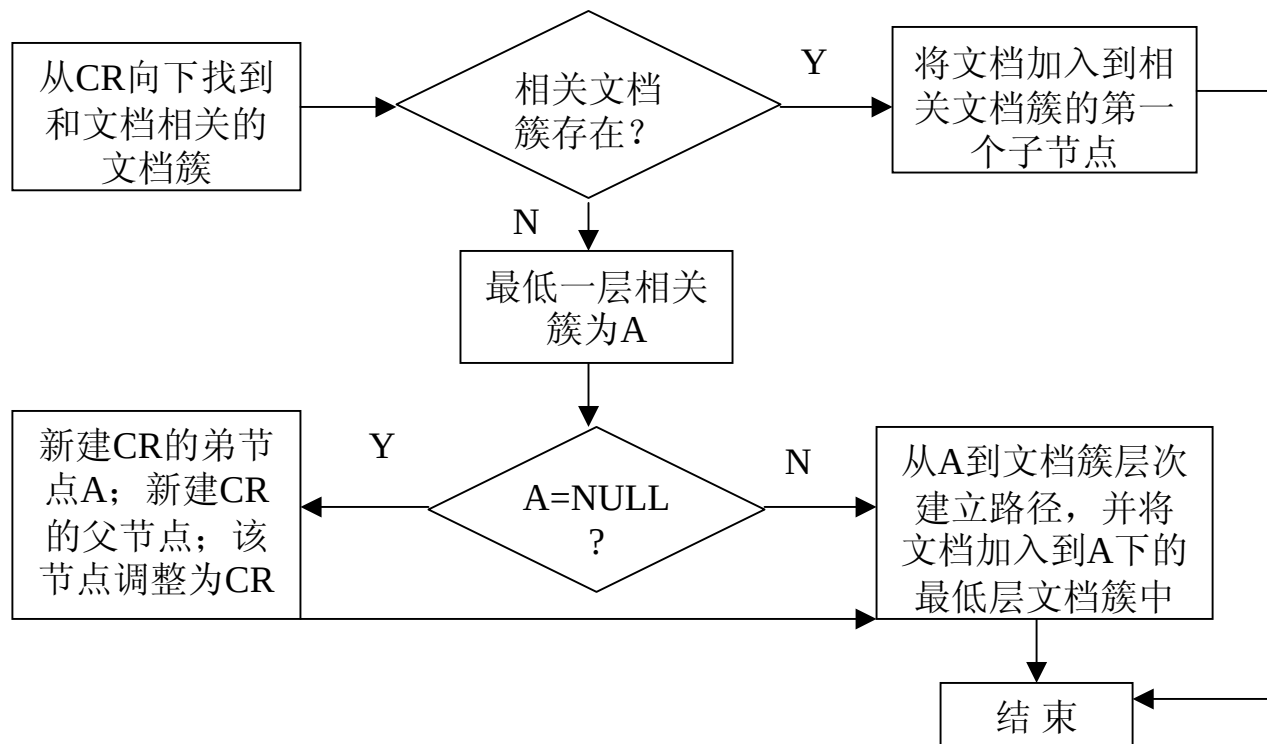
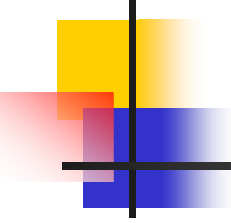


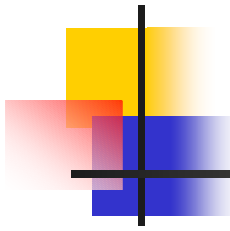
图3 插入聚类流程图



一种面向海量实时数据的信息检索算法

优化聚类:

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。
 - 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($H_C - 1$)层、第($H_C - 2$)层、...、第0层执行第1、2步的动作。



一种面向海量实时数据的信息检索算法

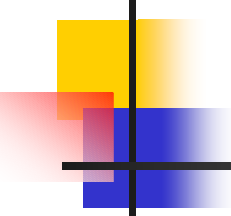
- 簇树属性
 - 节点数
 - 聚类半径
 - 质心
 - 簇树的高



一种面向海量实时数据的信息检索算法

■ 基于簇聚类的单链接算法

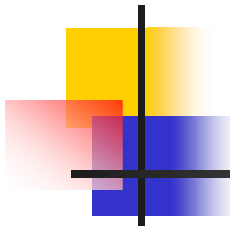
- 将群发邮件甄别模块放在簇聚类的两个聚类阶段之间。
- 完成插入聚类后，利用插入聚类的结果在文档簇内部执行群发邮件甄别，尽可能过滤掉垃圾邮件，再执行优化聚类。
- 过滤掉绝大多数的群发邮件，可以大大减轻优化聚类的压力，从而进一步增加数据处理能力。



一种面向海量实时数据的信息检索算法

■ 算法性能测试结果

- 以 CERNET 华（东）北地区的海量实时网络数据环境为依托
- 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。即：该小时内的群发邮件绝大多数可以被甄别出来。
- 群发邮件甄别算法的过滤率超过**95%**。



一种面向海量实时数据的信息检索算法

■ 小 结

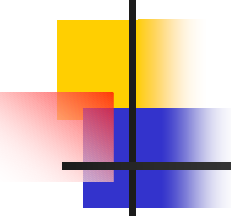
- 针对海量实时网络数据的信息检索展开研究，引入数据挖掘中的簇聚类技术，同时结合向量模型，提出两段向量簇聚类算法。
- 算法通过插入聚类完成对实时数据的高效收集和初级聚类，通过第二阶段的优化聚类解决多维数据聚类过程中所产生的质心漂移现象，提高对海量网络数据的实时处理能力和查询性能。
- 将单链接算法与簇聚类结构相结合，强化垃圾邮件的过滤，提高群发邮件的甄别效率，节约存储空间并进一步提高信息处理效率。



一种面向海量实时数据的信息检索算法

■ 未来工作方向

- 簇树聚类算法本身，尤其是优化聚类算法的设计。
- 跨簇查询的研究。



一种面向海量实时数据的信息检索算法

感谢聆听！