学校编码：10384　　分类号_____密级_____
学号：24320121152278　　UDC_____

厦门大学

硕 士 学 位 论 文

基于 Hadoop 的短文本聚类算法的研究与应用

Research and Application of Short Text Clustering Algorithms Based on Hadoop

王志沿

指导教师姓名：王 备 战 教 授
专 业 名 称：计算机软件与理论
论文提交日期：2015 年 4 月
论文答辩日期：2015 年 5 月
学位授予日期：2015 年 月

指 导 教 师：_____

答辩委员会主席：_____

2015 年 5 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　月　日解密，解密后适用上述授权。

（　√　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘 要

自从互联网开始普及，人们就身处在一个信息爆炸的时代，人们对待生活、工作的思维方式开始逐渐在改变。在 Web2.0 的 UGC（User Generated Content）时代，社交网络平台作为互联网发展的一个重要分支，成为了人们很重要的沟通、交流和营销的公开平台。社交网络平台上每天产生的数据是海量的，如何运用好这些数据宝藏，成为了一个热门的研究课题。

在数据分析方面，传统的统计抽样方法在面临海量的快速增长的数据时显得过时和力不从心，利用全体数据而不是部分抽样的数据成为了新的研究方法。为了达成该目的，仅依靠硬件的更新提速来提高机器的运算能力是无法完成的。因此，如何巧妙地运用云计算等弹性计算架构成为了人们关注的问题。社交网站作为 UGC 时代的支柱领域，每天都有海量的数据产生，如果能运用好这些数据，将是一笔巨大的财富。

论文以目前新浪微博平台为研究对象，针对其在文本聚类和话题文本推荐上的不足，研究了文本聚类算法和分布式技术，改进了聚类算法和相似度计算公式，实现一个基于分布式的短文本聚类，并将聚类的结果根据用户的输入进行文本推荐的应用。论文的主要工作如下：

首先，研究 Hadoop 平台下的 HDFS、MapRedue 和 HBase 三大基于 Google 核心技术实现的开源项目。包括 Hadoop 平台的优点、HDFS 的读写流程、MapReduce 的编程模型和 HBase 的结构。

其次，阐述利用网页爬虫与微博 API 两种不同的微博数据抓取方式的原理并对比其优缺点，数据的预处理方法以及根据特征权重表示为向量空间模型的方法，介绍了相似度计算方法及其改进、K-means 聚类算法和 Single-Pass 聚类算法的原理以及聚类算法选择，并对 Single-Pass 算法进行改进，设计了一个测试实验验证改进后的聚类算法和相似度计算方法的改进效果。

最后，在 Hadoop 平台下，使用改进后的 Single-Pass 聚类算法和相似度计算方法，对抓取的海量微博文本进行分布式聚类，并对用户输入的微博文本进行相似的微博推荐。

实验表明，论文使用的技术方法是有效可行的，可以较为准确地识别出微博

文本中的关键话题进行相似的微博文本推荐，且对比新浪微博平台自带的搜索工具后，发现微博平台的搜索工具无法完成相同的功能，因此论文使用的方法技术具有一定的实用性、新颖性。

关键词：Hadoop；短文本聚类；文本推荐

Abstract

Ever since the Internet began to come into our daily life, people are living in an era of information explosion. People's ways of deal with life and work begin to change. In the UGC (User Generated Content) Web2.0 era, social networking platforms become an important branch of Internet development, being a very important open platform for people's communication, messages exchanging and marketing. Data on social networking platforms are mass produced every day, how to make good use of these data treasures, has become a hot research topic.

As for data analysis, the traditional statistical sampling methods become obsolete and inadequate when facing massive and rapidly growing data. The use of all the data, rather than part of the sampling data has become a new research method. To achieve this purpose, only relying on updated hardware acceleration to improve the machine's computing power is not complete. Therefore, how to cleverly make use of cloud computing and other elastic computing architecture has become an issue of concern. Social networking sites as pillars of UGC era are generating vast amounts of data every day, if we can make good use of these data, it will be a great asset.

Aiming at the current shortage of Sina microblogging platform on the topic of short text clustering and text recommending, this dissertation study the text clustering algorithm and distributed technologies, improved clustering algorithms and similarity calculation formula, developed an application based on distributed short text clustering technologies which can recommend clustering result text according to the user input. The main work is as follows:

Firstly, this dissertation study three core open source projects HDFS, MapRedue and HBase in Hadoop platform based on Google technology. Including the advantages of Hadoop platform, HDFS reading and writing processes, MapReduce programming model and the structure of the HBase.

Secondly, this dissertation introduce the principles of the elaborate use of web crawler and Weibo microblogging Data API in fetching data and compare their ad-

vantages and disadvantages, as well as data preprocessing method based on feature weight expressed as a vector space model approach. This dissertation describes the similarity calculation formula and its improvement, principle of K-means clustering algorithm and Single-Pass clustering algorithm and algorithm choosing, and the Single-Pass algorithm improvements. This dissertation Design a testing experiment to verify the effect of similarity calculation formula improvement and clustering improvement.

Finally, in the Hadoop platform, this dissertation uses the improved Single-Pass clustering method and similarity calculation formula, distributed clusters the massive microblog text fetched, recommends similar clustering result text according to the user input.

Experiments show that the technical methods used in this dissertation is feasible and effective, and can be more accurate in identifying and recommending similar micro-blog text, and after comparing search tools in Sina microblog platform we can find that microblog platform search tool can not perform the same function, so the method used in this dissertation is practical and novel.

Keywords: Hadoop; Short Text Clustering; Text Recommendation

目 录

Contents

第一章 绪论

1.1 研究背景

近年来，移动互联网的浪潮来袭，随着移动设备的增加和网络基础设施带宽等的优化，人们开始进入移动互联网时代。互联网上的数据再一次迎来了高速的爆炸式增长，据 2015 年 02 月发布最新的《互联网络发展状况统计报告》[1]显示，我国使用手机上网的网民规模已达到 5.57 亿，总网民规模在 6.49 亿（2014 年 12 月底数据）。使用手机上网人数比例已经在 2014 年 6 月份的统计[2]中超过传统 PC 上网。我国互联网的普及率为 47.9%，相比较 2013 年底，提升了 2.1%，我国的网民规模和互联网普及率随着时间的变化如图 1-1 所示。

图 1-1：中国网民规模和互联网普及率

该报告还指出，截至 2014 年 12 月，我国的微博用户规模已经达到 2.49 亿，在网民使用率为 38.4%。其中，使用手机微博的用户数为 1.71 亿，使用率为 30.7%。2013-2014 年微博用户规模及网民使用率如图 1-2 所示。

可以看到微博的总用户数在减少，然而，自从 2014 年 8 月左右，腾讯解散旗下微博事业部，只维持腾讯微博的基本运营，网易和搜狐等公司也纷纷减少对微博的投入。各个微博客服务商之间竞争逐步趋缓，用户群体主要向新浪微博倾斜，这也促使新浪“微博”用户也较以往略有提升，微博客一家独大的格局明朗。因此，在 2014 年，新浪微博在其一家独大后，赴美上市。

社交媒体[3]与社交类沟通应用体现出不同的应用属性。2014 年上半年的“马航事件”和 2014 年下半年的“冰桶挑战”凸显了新浪微博作为社交媒体的快速的传播速度、深远的传播范围和积极的社会影响力。

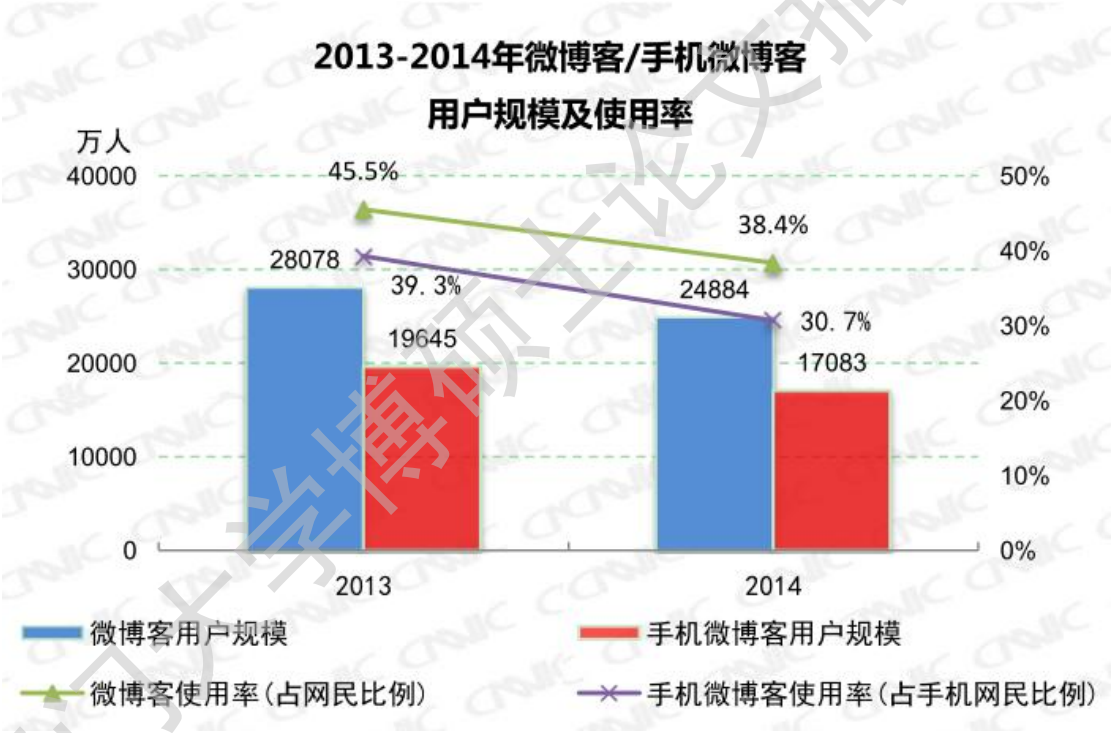

图 1-2：2013-2014 年微博用户规模及网民使用率

新浪微博的蓬勃发展，在吸引了数以亿计的用户的同时，产生了海量的微博文本数据以及相关的海量数据处理需求。

在数据分析方面，传统的统计抽样方法在面临海量的快速增长的数据时显得过时和力不从心，利用全体数据而不是部分抽样[4]的数据成为了新的研究方法。为了达成该目的，仅依靠硬件的更新提速来提高机器的运算能力是无法完成的。因此，如何巧妙地运用云计算等弹性计算架构成为了人们关注的问题。社交网站

作为 UGC 时代的支柱领域，每天都有海量的数据产生，如果能运用好这些数据，将是一笔巨大的财富。

对于技术架构来说，由于基础硬件及带宽的支撑，互联网的传输数据及存储数据也急剧增长，海量数据处理便成了一个十分重要的话题。但幸运的是，现在的互联网不仅仅在硬件设施上得到了很大的提高，而且在软件架构等方面也得到了十分可观的提高，对于海量数据处理的技术，也提出了诸多很有效果的模型与应用。例如 Google 提出的 BigTable、MapReduce、GFS（Google File System）[5][6][7]等技术；再如云计算、云框架、网格分布式计算等，极大地满足了互联网的海量数据处理的发展与需求。

1.2 研究内容及现状

在大数据时代背景下，微博成为了人们重要的沟通、营销工具[8]，这些行为产生了海量的微博数据。这些海量数据的处理和利用成为一个热门的研究课题，而对这些海量数据最为直观且需求量最大的应用之一，就是基于微博数据的文本推荐。

要做到文本推荐，首先涉及到的技术有文本聚类和分布式技术下面介绍这些技术的研究现状综述。

1.2.1 文本聚类

聚类[9]是一种无监督的、在没有给定“标准答案”的情况下，利用算法自身对数据的认识，自动地让数据聚集分割为一个个子集的过程，分割后的每个子集被称为簇或者组。通过事先定义好的相似度计算方法，如定义好的数据间距离的计算，可以让组内的数据满足相似度较高的条件。聚类是数据挖掘[10]领域中一项基本的数据分析技术。

同样的，文本聚类[11]（Text clustering）主要也是一种无监督的机器学习方法，可以自动的根据相似度的定义把文本归类。比较经典的聚类算法有 K-Means[12]算法和 Single-Pass[13]算法。

K-Means 算法的基本流程是：首先在数据集 $D=(d_1,d_2,...,d_n)$ 中随机设定 K