DOI: 10.16353/j.cnki.1000-7490.2015.03.027

● 杨　京，王效岳，白如江，祝　娜（山东理工大学　科技信息研究所，山东　淄博　255049）

大数据背景下数据科学分析工具现状及发展趋势*

摘　要：文章根据大数据时代的特征，分析了海量数据给数据科学分析工具带来的主要挑战，介绍了为应对挑战而发展的大数据分析工具，并对比分析了R语言、Rapid Miner、Mahout三种数据科学中比较流行的大数据分析工具，发现R语言和Rapid Miner功能全面，而Mahout具有突出的大数据分析能力，最后指出了数据科学分析工具的发展趋势。

关键词：数据科学；R语言；大数据

Abstract: According to the features of big data era, this paper analyzes the main challenges that massive data bring to the analysis tool of data science. The paper introduces the big data analysis tool in response to challenges. Then, the paper carries on the comparative analysis of R language, Rapid Miner and Mahout 3 popular analysis tools of big data in data science, which finds that R language and Rapid Miner have fully functions and the Mahout has more outstanding analysis capability of big data. Finally, the paper points out the development trend of data science analysis tool.

Keywords: data science; R language; big data

数据爆炸把人类带入了一个大数据时代。大数据的火爆，带动了政府、国内学术界和产业界等对大数据的热情。“Nature”和“Science”等国际顶级学术刊物相继出版专刊来探讨大数据。2008年“Nature”推出专刊“Big Data”，从互联网技术、环境科学、网络经济学等多方面介绍了海量数据带来的挑战[1]；2011年“Science”出版数据处理的专刊“Dealing with Data”，探讨了数据洪流带来的挑战[2]。政府部门同样高度关注。2012年3月，美国公布了“大数据研发计划”[3]；欧盟在过去几年对科学数据基础设施建设投资了1亿多欧元，并将数据信息化基础设施建设作为Horizon 2020计划的优先项目之一[4]。

大数据的热潮激发了科研人员开始考虑数据科学问题。2014年2月“科学数据大会”在北京召开，大会以“科研大数据与数据科学”为主题，研讨了大数据时代科研数据管理、共享与应用的新趋势，以及科研大数据面临的关键问题和挑战，探索数据科学的科学内涵与发展方向[5]。

数据科学是一门横跨信息科学、网络科学、经济学等诸多领域的新兴交叉学科，依然处于发展初期。本文分析了数据科学这门学科的发展现状，在此基础上介绍了数据爆炸给数据科学中的数据分析工具带来的挑战以及能够应对挑战的大数据分析工具。然而并不是所有的工具都具备全面的功能，它们各具特点和优缺点，随后选取了R语言、Rapid Miner、Apache Mahout三种主流的大数据分析工具，概述了工具特点，并以表格的形式对其在大数据处理能力、可视化等方面进行了分析。

1　数据科学

大数据的热潮，催生了一门新的学科即数据科学。数据科学正处于发展初期，是一门不断发展的学科。数据科学的核心涉及用自动化的方法来分析海量数据，并从中提取知识[6]。在几乎所有的知识发现领域，数据科学提供了一种强大的新方法来探索发现，它为拥有大量数据但不知怎样从数据中提取价值的公司提供了一种新的见解来源。伴随着这种自动化方法的发展，数据科学正在帮助创造新的科学分支并影响着社会科学和人文科学领域。

数据科学融合了多门学科并且建立在这些学科的理论和技术之上，包括数学、概率模型、统计学、机器学习、数据仓库、可视化等[7]（如图1所示）。在实际应用中，数据科学包括数据的收集、清洗、分析、可视化以及数据应用整个迭代过程，最终帮助组织制定正确的发展决策。数据科学的从业者称为数据科学家[8]。数据科学家是有着开阔视野的复合型人才，他们既有坚实的数据科学基础，如数学、统计学、计算机学等，又具备广泛的业务知识和经验。数据科学家通过精深的技术和专业知识在某些科学学科领域解决复杂的数据问题，从而制定出适合不同决策人员的大数据计划和策略，他们被认为是21世纪“最性感”的职场人才。

* 本文为山东省自然科学基金项目“大规模学术文献并行处理与语义分类研究”（项目编号：ZR2011GL025）和山东理工大学人文社会科学发展基金资助项目的成果之一。

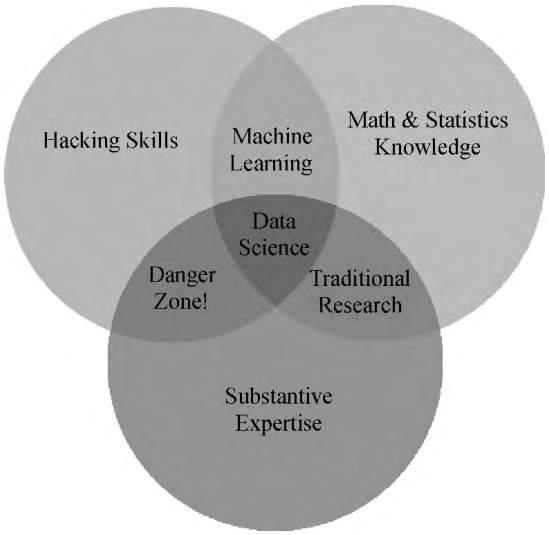

图1　数据科学

数据科学目前还没有明确的基础理论，人们对数据科学的定义各不相同。许多学者立足各自的视角对数据科学的基础理论提出了不同的观点，例如 V. Dhar 将数据科学定义为研究从数据中提取知识的一门学科[8]。J. Leak 认为数据科学其关键词是“科学”而不是“数据”[9]。复旦大学数据科学研究中心的朱扬勇教授则认为数据科学是关于数据的科学或者研究数据的科学，定义为：研究探索 Cyberspace 中数据界奥秘的理论、方法和技术，研究的对象是数据界中的数据[10]。因此，数据科学要作为一门独立的学科存在，还需要更多的学术认同和大量长期的实践积累。

2　数据科学分析工具面临的主要挑战

大数据作为数据科学中非常重要的一方面，在给科学和教育事业的发展提供了巨大机会的同时，也给前沿科学项目带来了极大的挑战。正如 J. Gray 所说的那样，数据洪流，尤其是科技领域的数据洪流，已经发出了构建新的研究基础设施的挑战[11]。数据爆炸主要给数据科学分析工具带来以下 3 方面挑战。

2.1　数据格式的多样性

进入大数据时代，如今的数据量正在急剧增长，数据的格式也是多样化。比如银行、超市的数据会是文本格式，YouTube 中的数据是图像视频格式，数字电话的数据是语音格式等。数据形式除了传统的关系型数据外，还包括来自网页、电子邮件、社交媒体论坛、搜索索引、互联网日志文件等原始、非结构化和半结构化的数据。

所以，面对数据量如此之大且形式多样的数据，要求数据分析工具能够把处理结构化数据的方法和新的非结构化数据的方法有机结合。

2.2　传统数据算法的失效

进行数据分析时，需要用于数据挖掘和分类聚类的更好的算法。然而，聚类算法不是对数曲线式（$N \log N$）或线性（顺序 N）的规模，而是典型的 N 立方规模，当 N 变得很大时，一些方法就会失效[12]。因此，在面对海量的数据处理时，许多传统的算法会失去有效性。于是需要发明新的算法并且要求算法具有良好的伸缩性，能够应对 PB 级别的数据。

另外，大数据的许多应用具有实时性的特点，这种情况下大数据应用最主要的指标就不再是算法的准确率，算法需要在准确率和实时性之间取得一个平衡[13]，如在线的机器学习算法。

2.3　超大规模的数据可视化

可视化是解释海量数据最有效的手段之一。通过交互界面的支持来进行可视化分析，不仅能监测和验证预计，还能发现没有预见到的内容。有效的可视化工具与数据分析一样，是以数学根基和强健算法为基础的。

然而海量数据给可视化技术带来了诸多挑战，北京大学的袁晓如研究员在“科学数据大会”上发表《大数据可视分析》报告，指出了大数据可视化主要面对融合不同尺度的多元异构数据、任务复杂度的可扩展性以及交互扩展性等挑战[5]。此外还有如原位分析、算法、并行化、数据移动、不确定性的量化、传输和网络架构等一系列挑战[14]。因此，我们需要促进可视化技术的重大进展，来支持从巨大且复杂的数据集中提取含义。

3　数据科学分析工具

面对数据爆炸式增长给科学项目带来的极大挑战，需要更加善于开发相关的技术和工具，支撑从数据采集、数据管理到数据分析以及数据可视化整个数据处理周期。经过研究人员的不懈努力，开发的技术和工具不断推陈出新。在数据存储方面，以 Google 为代表的公司分别开发了自己的 NoSQL 数据库，比如 Google 公司的 BigTable[15]，VMware 公司的 Redis[16]，Microsoft 公司的 Azure Tables[17] 等，成功解决了不同格式数据的存储和管理问题。

数据分析方面，Google 发明了大数据分析的分布式编程模型 MapReduce[18] 技术，实现了并行计算。随之发展而来的数据科学分析工具更是形式多样，最为流行的当属 Yahoo 的开源项目 Hadoop[19]，以分布式文件系统（HDFS）和 MapReduce 为核心的 Hadoop 为用户提供了系统底层细节透明的分布式基础构架。除了 Hadoop 外，还有许多面向大数据分析的数据科学分析工具，如 HPCC、R 语言、Storm、Apache Drill、Rapid Miner、Mahout 等，

它们有的是专门针对特定的大数据分析应用，有的是完整的分析平台。

3.1 主要数据分析工具

随着这些数据科学分析工具的发展，它们一方面成功解决了数据科学中的算法失效、超大规模数据可视化等一系列挑战；另一方面各具特点和优缺点。例如 Mahout 具有优秀的大数据处理能力，不仅处理数据量大且速度快，但可视化能力差。

接下来选取 R 语言、Rapid Miner、Mahout 三种主流的数据科学分析工具，对其概述并以表格的形式对三者的主要特点进行了比较分析，工具基本情况如下。

1) R 语言[20]是一种用于统计计算和作图的编程语言和环境，采用命令行工作方式，在 GNU 协议下免费发行，其源代码可供自由下载和使用。R 的网站 CRAN 上提供了大量的第三方程序包，内容涵盖了经济学、社会学、统计学、生物信息学等诸多方面，这也是为什么越来越多的各行各业的人员喜爱 R 的一个重要原因。

针对传统分析软件的扩展性差以及 Hadoop 的分析功能薄弱的弱势，研究人员致力于将 R 语言和 Hadoop 的集成。R 作为开源的统计分析软件，通过 R 与 Hadoop 的深度集成，把数据计算推向并行处理，使 Hadoop 获得强大的深度分析能力。

2) Rapid Miner[21]原名 Yale，是一种用于数据挖掘、机器学习以及商业预测分析的开源计算环境。其既可以使用简单的脚本语言进行大规模进程操作，也可以通过 Java API 或 GUI 模式进行操作。因为其具备 GUI 特性，所以对于数据挖掘的初学者比较容易入门。

Rapid Miner 6 具有友好而强大的工具箱，提供快而稳定的分析，可以在短时间内设计好一个原型，使得数据挖掘过程中的关键决策尽可能早地实现。帮助减少客户流失、进行情感分析、预测性维护以及市场直销等。

3) Apache Mahout[22]起源于 2008 年，其主要目标是构建一个可伸缩的机器学习算法的资源库，它提供了一些经典的机器学习算法，旨在帮助开发人员更加方便快捷地创建智能应用程序。目前，Mahout 的项目包括频繁子项挖掘、分类、聚类、推荐引擎（协同过滤）。

表 1 工具比较分析

指标	R 语言	Rapid Miner	Mahout
支持平台	能够在不同体系结构的计算机系统上运行，例如 UNIX（包括 Linux 和 FreeBSD）、MacOS 和 Windows	能够在所有主要平台和操作系统上运行	能够在 Linux、Windows 以及 Mac OSX 等平台安装与配置，但最好的支持系统为 Linux
模型算法	支持大多数挖掘算法。例如 Naïve Bayes、K-Means、EM、Neural Network、SVM、Apriori、KNN 等	同 R 语言一样，支持大多数分类、聚类以及关联分析的模型算法	实现了部分数据挖掘算法。Mahout 支持的算法大多是距离的算法，许多的数据挖掘算法无法实现 MapReduce 并行化，这是 Mahout 的一个空白点
数据格式兼容性	兼容性好。支持各种格式的结构化、半结构化以及非结构化的数据	兼容性好。支持 Excel、Arff、SPSS、Dbase、CSV 等多种格式的数据源以及 PDF、ASCII、XML 和 HTML 格式的网页和文本文档等	兼容性差。Mahout 下处理的文件必须是 SequenceFile 格式的，所以需要把 Txtfile 转换成 SequenceFile
运行速度	运行速度稍慢	运行速度快。Rapid Miner6 可以在 5 分钟内甚至更少的时间使用程序应用模板开始进行预测	运行速度快，基于 Hadoop 的分布式处理之上
NoSQL 数据库	支持	支持	支持
二次开发	不支持	不支持	支持。但对技术要求性高。开发者不仅要对算法有一定的了解，同时要求具有编写符合 MapReduce 流程的伪代码算法以及把伪代码转换为实际代码的能力
可视化	可视化能力强。R 语言具有强大的图形用户界面，可以轻易地绘制出高质量的图片，包括饼图、点图、正态分布图、趋势图等，并且绘制出的图形可以输出 pdf、jpg、png 等各种格式，满足出版印刷的格式要求	可视化能力强。Rapid Miner 具有卓越的绘图功能。允许用户把数据任意转变成散点矩阵、图表以及在线 1D、2D、3D 图等，具有非常优秀的视觉效果	可视化能力差。Mahout 不具有绘图功能，只支持向量和矩阵表示
大数据处理	支持。通过 R 与 Hadoop 的深度集成，把计算推向数据并行处理。目前主要有两种方法[23]。第一种方法是，利用 Hadoop 的 MapReduce 分布式计算将 PB、TB 量级的数据缩小到 GB 量级，再加载到 R 语言中进行处理。第二种方法是，直接采用支持 Hadoop 的 R 包，利用 R 语言操作存放在 HDFS 中的数据，并利用 R 完成 MapReduce 算法，用来替代 Java 的 MapReduce 实现	支持。为了进行海量数据分析，在 Hadoop 基础上对 Rapid Miner 进行扩展，创建扩展接口 Radoop[24]。Radoop 为 RapidMiner 提供其他的操作接口，可以在 Hadoop 集群上运行任务进行大数据处理。并且，可以重用 Hive 和 Mahout 中的某些数据分析功能	支持。利用了 Hadoop 的并行计算能力。Mahout 的核心目标就是利用 Hadoop 的并行计算能力实现一系列机器学习算法。Mahout 的算法运行在 Apache Hadoop 的平台下，通过 Hadoop 的分布式计算 MapReduce 模式实现，使 Mahout 在处理海量数据时具有较快的速度，极大提升了算法可处理的数据量和处理性能

Mahout 目前支持两种根据贝氏统计来实现内容分类的方法。第一种是使用简单的支持 Map-Reduce 的 Naive Bayes 分类器。Naive Bayes 分类器以准确性高和速度快而著称，但其假设数据是完全独立的；第二种是 Complementary Naive Bayes，它纠正了 Naive Bayes 方法中的一些缺陷，同时维持了前者的简单性和速度。

3.2 特点分析

针对大数据时代给数据科学分析工具带来的挑战以及对工具性能的要求，本文选取了模型算法、可视化、大数据处理能力等一系列工具特点指标，以表格的形式对 3 种工具进行了分析和比较，见表 1。

表 1 中对 3 种工具的若干指标进行了比较分析。由表 1 可以看出，R 语言作为一种开源的编程语言和软件环境，具有比较全面的功能。R 语言支持大部分模型算法且支持不同格式、不同数据源的数据。R 不仅能够进行大数据集的分析并且具备卓越的绘图功能，使其对大数据集进行可视化，是一种广受欢迎的数据分析和可视化工具。

Rapid Miner 是一款易于使用的可视化预测分析软件，对普通用户而言容易入门。一方面 Rapid Miner 6 摆脱了旧版本在大数据处理方面的缺陷，通过与 Hadoop 的集成具有了良好的大数据分析能力；另一方面可视化能力优秀，具有 R 语言和 Mahout 不支持的 3D 图。

由于 Mahout 是基于 Hadoop 的数据挖掘和机器学习的算法框架，因此 Hadoop 的优点即 Mahout 的优点。Mahout 在大数据集分析上表现出明显的优势，充分利用了 MapReduce 的并行计算能力，其训练样本数量大，能够高效完成计算任务。但是弊端也十分明显，首先 Mahout 目前支持的算法较少，尽管算法一直在增加；其次，Mahout 可视化能力差。

4 发展趋势

通过以上对数据科学分析工具的介绍和比较以及大数据时代对工具特点的要求，本文认为，数据科学分析工具主要有以下几点发展趋势：

1）大数据集分析。大数据时代毫无疑问要求数据科学分析工具能够胜任海量数据的分析。其次，数据价值与数据容量和种类是密切相关的。一般说来，数据容量越大，种类越多，包含的信息量越大，挖掘的潜在价值也越大。为了实现全数据分析从而发掘新的并且有价值的洞察力，要求数据科学数据分析工具能够综合分析海量且格式多样的数据。

2）优秀的可视化能力。数据分析是数据处理的核心步骤，但是如果分析的结果正确而没有运用适当的方法解释，那么得到的结果会让用户难以理解。直观有效地展示分析结果，可以让人更容易地接受数据分析工具传达的关键信息。在大数据时代，数据量不仅大而且烦琐，帮助人们直观地发现数据中包含的信息和知识，可视化是最有效的途径之一。

3）数据分析以分布式为主。大数据时代，仅靠过去单一的数据分析工具已经不能胜任海量数据的分析，采用分布式架构来提高系统的扩展性已成为必然。毫无疑问，Hadoop 已经成为当今大数据处理领域的王者技术。分布式处理技术极大地提高了数据分析的效率和速度，未来如 Mahout 等的分布式大数据处理工具将替代传统工具，占据主要地位。

5 结束语

大数据时代，倘若能够更加有效地组织和使用数据，人们将得到更多的机会发挥科学技术对社会发展的巨大推动作用。因此，随之要求数据分析工具的不断发展，能否高效、准确地挖掘出数据中蕴含的潜在价值，是衡量数据分析工具的价值之处，也是数据科学的关键所在。

而大数据的浪潮促进了“数据科学”逐步发展成一门独立的学科。当前，数据科学的定义还没有明确，但假以时日，数据科学将成为一门专门的学科，拥有完善理论基础和学科技术，被越来越多的人所认知。各大高校也将设立专门的数据科学类专业，催生一批与之相关的新的就业岗位。未来几年，数据科学家必定成为各行各业的紧缺人才。□

参考文献

[1] Big Data-Nature [EB/OL]. [2014-04-10]. http://www.nature.com/.

[2] Dealing with Data-Science [EB/OL]. [2014-04-10]. http://www.sciencemag.org/.

[3] 美国政府出台大数据研发计划 [DB/OL]. [2014-04-10]. http://www.most.gov.cn/.

[4] 欧盟 Horizon 2020 规划科研基础设施的发展 [EB/OL]. [2014-04-10]. http://eu.mofcom.gov.cn/.

[5] 2014 科学数据大会 [DB/OL]. [2014-04-10]. http://dc2014.codata.cn/.

[6] Date Science at NYU [EB/OL]. [2014-04-10]. http://datascience.nyu.edu/.

[7] Wikipedia Date Science [EB/OL]. [2014-04-10]. http://en.wikipedia.org/wiki/Data_science.

[8] DHAR V. Data science and prediction [EB/OL]. [2014-04-10]. http://cacm.acm.org/magazines/2013/12/169933-data-science-and-prediction/fulltext.

[9] LEAK J. The key word in “Data Science” is not data, it is science. [EB/OL]. [2014-04-20]. http://simplystatistics.org/2013/12/12/the-key-word-in-data-science-is-not-data-it-is-science/.

（下转第 144 页）

ring intention under self-efficacy, trust, reciprocity, and shared-language [J]. Computers & Education, 2013, 68: 223-232.

[27] CHIU C M, HSU M H, WANG E T G. Understanding knowledge sharing in virtual communities: an integration of social capital and social cognitive theories [J]. Decision Support Systems, 2006, 42 (3): 1872-1888.

[28] 张鼐，周年喜. 社会资本和个人动机对虚拟社区知识共享影响的研究 [J]. 情报理论与实践，2012，35 (7): 56-60.

[29] 雷雪，焦玉英，陆泉，等. 基于社会认知论的 Wiki 社区知识共享行为研究 [J]. 现代图书情报技术，2008 (2): 30-34.

[30] 黄照贵，颜郁人. 以关系承诺观点探讨虚拟社群不同参与程度成员之行为 [J]. 信息管理学报，2009，16: 57-81.

[31] 姜洪涛. 虚拟社区组织公民行为影响因素研究 [D]. 重庆: 重庆大学，2009.

[32] 石密，李旭培，时勘. 从心理安全的视角看虚拟社区的知识共享 [J]. 人类工效学，2012，18 (1): 74-76.

[33] ZHANG Y X, FANG Y L, WEI K K, et al. Exploring the role of psychological safety in promoting the intention to continue sharing knowledge in virtual communities [J]. International Journal of Information Management, 2010, 30 (5): 425-436.

[34] ZHA X J, LI J, YAN Y Y. Understanding preprint sharing on sciencepaper online from the perspectives of motivation and trust [J]. Information Development, 2013, 29 (1): 81-95.

[35] 王伟军，甘春梅，刘蕤. 学术博客知识交流与共享心理诱因的实证研究 [J]. 情报学报，2012，31 (10): 1026-1033.

[36] 徐美凤. 基于 CAS 的学术虚拟社区知识共享研究 [D]. 南京: 南京大学，2011.

[37] PAROUTIS S, SALEH A A. Determinants of knowledge sharing using Web 2.0 technologies [J]. Journal of Knowledge Management, 2009, 13 (4) 52-63.

[38] SUH A, SHIN K S. Exploring the effects of online social ties on knowledge sharing: a comparative analysis of collocated vs dispersed teams, [J]. Journal of Information Science, 2010, 36 (4): 443-463.

[39] CHI L, CHAN W K, SEOW G, et al. Transplanting social capital to the online world: insights from two experimental studies [J]. Journal of Organizational Computing and Electronic Commerce, 2009, 19 (3): 214-236.

[40] 王东. 虚拟学术社区知识共享实现机制研究 [D]. 长春: 吉林大学，2010.

[41] 王健. 虚拟学术社区中知识共享行为的博弈分析 [D]. 武汉: 华中师范大学，2013.

作者简介: 张敏，女，1974 年生，博士，副教授。研究方向: 信息服务与用户。通讯作者。
郑伟伟，女，1988 年生，硕士生。

收稿日期: 2014 - 09 - 01

(上接第 137 页)

[10] 朱扬勇，熊赟. 数据学与数据科学 [EB/OL]. [2014-04-20]. http://www.dataology.fudan.edu.cn.

[11] GRAY J. Jim Gray on eScience: atransformed scientific method [R]. The FourthParadigm: Data-intensive Scientific Discovery, 2009.

[12] HEY T. 第四范式: 数据密集型科学发现 [M]. 北京: 科学出版社，2012.

[13] 孟小峰，慈祥. 大数据管理: 概念、技术与挑战 [J]. 计算机研究与发展，2013，50 (1): 146-169.

[14] WONG P C, SHEN H-W, JOHNSON C R, et al. The top 10 challenges in extreme-scale visual analytics [J]. Computer Graphics and Applications, 2012, 32 (4): 63-67.

[15] CHANG F, DEAN J, GHEMAWAT S, et al. Bigtable: a distributed storage system for structured data [J]. ACM Transactions on Computer Systems (TOCS), 2008, 26 (2): 4.

[16] Redis [EB/OL]. [2014-05-10]. http://redis.io/.

[17] Azure Tables [EB/OL]. [2014-05-10]. http://azure.microsoft.com/.

[18] DEAN J, GHEMAWAT S. MapReduce: simplified data processing on large clusters [J]. Communications of the ACM, 2008, 51 (1): 107-113.

[19] Hadoop [EB/OL]. [2014-05-20]. http://hadoop.apache.org/.

[20] R 语言 [EB/OL]. [2014-05-20]. http://www.r-project.org/.

[21] Rapid-I [EB/OL]. [2014-04-20]. http://rapid-i.com/content/view/181/196/.

[22] Mahout [EB/OL]. [2014-04-21]. https://mahout.apache.org/.

[23] 杨霞，吴东伟. R 语言在大数据处理中的应用 [J]. 科技资讯，2013 (23): 19-20.

[24] PREKOPSÁK Z, MAKRAI G, HENK T, et al. Radoop: analyzing big data with rapidminer and hadoop [C] //Proceedings of the 2nd RapidMiner Community Meeting and Conference (RCOMM 2011), 2011: 1-12.

作者简介: 杨京，男，1990 年生，硕士生。
王效岳，男，1961 年生，博士，教授。
白如江，男，1979 年生，讲师。
祝娜，女，1988 年生，硕士生。

收稿日期: 2014 - 09 - 15