密级：　　保密期限：

北京郵電大學

硕士学位论文

题目：基于电商大数据的推荐算法研究

学　　号：2014140099

姓　　名：卢嘉颖

专　　业：电子与通信工程

导　　师：李勇

学　　院：信息与通信工程学院

二〇一六年六月

独创性（或创新性）声明

本人声明所呈交的论文是本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京邮电大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

申请学位论文与资料若有不实之处，本人承担一切相关责任。

本人签名：_____ 日期：_____

关于论文使用授权的说明

学位论文作者完全了解北京邮电大学有关保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属北京邮电大学。学校有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许学位论文被查阅和借阅；学校可以公布学位论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存、汇编学位论文。（保密的学位论文在解密后遵守此规定）

本学位论文不属于保密范围，适用本授权书。

本人签名：_____ 日期：_____

导师签名：_____ 日期：_____

基于电商大数据的推荐算法研究

摘 要

本课题主要针对电子商务产生的大数据进行研究，采用分布式系统存储并处理这些数据量大、多变、速度快且价值密度低的数据，比较协同过滤、逻辑回归、随即森林、GBDT 等多种机器学习推荐算法的准确率和召回率，并应用于分布式系统上，分析各算法的优劣，最终提出有创新性和适应于海量数据的算法。

关键词：推荐系统　机器学习　分布式系统　spark

ABSTRACT

KEY WORDS:

目 录

符号对照表

$(\cdot)^*$	复共轭
$(\cdot)^{\mathrm{T}}$	矩阵转置
$(\cdot)^{\mathrm{H}}$	矩阵共轭转置
$\mathbf{X}$	矩阵或向量
$\mathcal{A}$	集合
$\mathcal{A} \times \mathcal{B}$	集合 $\mathcal{A}$ 与集合 $\mathcal{B}$ 的 Cartesian 积，即 $\mathcal{A} \times \mathcal{B} = \{(a, b) : a \in \mathcal{A}, b \in \mathcal{B}\}$

第一章 绪论

随着移动互联网、物联网、云计算等新兴信息技术在社会各个领域的广泛应用，全球数据量正呈现出前所未有的指数型增长态势。与此同时，数据类型的丰富性及来源的多样性、数据产生的高速性与分析的实时性、数据的低价值密度等复杂特征日益凸显，标志着“大数据”时代的到来。大数据同云计算、物联网一样，是信息技术领域的重大技术变革。大数据的产生在很大程度上降低了消费者和企业之间的信息不对称程度。一方面，企业通过多元化的信息获取渠道掌握消费者的全面信息，提供的产品和服务更具针对性；另一方面，分散孤立的消费者同样通过多种渠道了解产品的各种信息，需求逐步呈现出个性化和多样化趋势。交易双方信息的愈加透明促进消费者与生产企业之间更加互动，消费者的个性化需求成为生产企业关注的核心。因此，大数据等新一代信息技术的发展使得消费者的地位日益重要，推动电子商务的价值创造方式发生转变，生产企业以消费者为中心创造高度差异化的产品和服务，并且引导消费者参与产品生产和价值创造。通过对海量和复杂的数据进行收集、整理与分析，不仅能够提升对社会经济发展的预测能力，而且能够不断地在各领域创新商业模式。本课题将分析大数据背景下的电子商务推荐算法的创新及其在大数据应用面临的挑战。

研究现状和发展趋势：当前，大数据已深耕于经济领域并创造了巨大的经济价值，美国将大数据上升为国家战略，英国开展了“数据权”运动，欧盟提出了开放数据战略，而中国也发布了大数据标准化白皮书。可以说，世界各主要经济体都将大数据视作未来国家竞争力的重要组成部分。在电子商务领域，大数据技术的发展给很多企业带来了广阔的发展机会。传统电子商务创新主要局限在电子商务的效率、便利化、营销方式等方面，大数据技术的广泛应用给电子商务的模式创新带来机遇。基于大数据的电子商务创新主要在于提炼大数据的价值并将其应用于电子商务的各个流程，形成新的商业模式[1]。其中，由推荐算法延伸出的推荐系统满足了大数据时代的消费者的个性需求，使得按需定制变为可能，得以创造实时化、差异化的产品及服务以满足各种长尾群体的需求。

北京邮电大学北京邮电大学（Beijing University of Posts and Telecommunications，BUPT）研究生院培养与学位办公室于 2014 年 11 月颁布了最新的《北京邮电大学关于研究生学位论文格式的统一要求》（下简称“要求”）[2]，对原有研究生学位论文的

格式要求做出了新的修订。但是迄今为止，研究生院尚未发布统一的论文模板。对于已经、正在或者即将撰写学位论文的同学都只能按照该要求的规定自行调整其学位论文的格式，一方面给大家增加了繁重的排版工作，另一方面也不利于统一全校的论文格式。

2007 年 9 月，北京邮电大学无线新技术研究所无线新技术研究所（Wireless Technology Innovation Institute，WTI）的王旭博士制作并发布了 latex-bupt——北京邮电大学博士毕业论文 LaTeX 模板（非官方版）[3]。该模板可以满足旧版官方论文格式要求[4]，但是在一些细节上的处理还有待改进，例如：

- 参考文献不能分列在各章末尾；
- 不能利用 BiBTeX 处理发表学术论文列表；
- 参考文献的格式上赏不能完全满足学校要求等。

2009 年，张煜博士发布了 buptthesis——北京邮电大学研究生学位论文 LaTeX 文档类（非官方版）[5]。该模板解决了 latex-bupt 中存在的问题，并且同样可以满足旧版官方论文格式要求[4]，但是仍然存在以下一些问题可以改进：

- 论文格式与最新版的官方论文格式要求[2] 有细微出入；
- 中文解决方案采用旧式 CJK 宏包，需要用户自行生成字体；
- 缺乏详细的用户使用文档，用户撰写论文过程中遇到的问题基本都需要登陆北邮人论坛发问，由张博逐一解答。

本模板在 buptthesis[5] 的基础上，增加了 XeTeX 编译引擎，使用 xeCJK 宏包作为中文解决方案。同时，本模板还根据北京邮电大学发布的最新的论文格式要求 [2] 进行模板格式的修改。本模板还提供了较为细致的用户使用文档，可以帮助初级用户快速上手使用本模板。

1.1 中文信息处理软件的国内外发展现状

中文信息处理软件可以分为字处理软件和排版软件两大类。字处理软件包括以下功能：字体、字号设定，英文断字，拼写和语法检查等。通常字处理软件处理文档的规模比较小，一般是作为办公自动化套件的一个重要组成部分，目前广泛使用

的中文字处理软件主要包括微软 Office 套件中的 Word、金山公司的 WPS，以及开源社区的 OpenOffice 等。排版软件则是针对大规模专业出版印刷而设计的一类软件，其主要功能是文字图像定位，基本图形绘制等。排版软件相对于字处理软件其专业针对性更强，目前广泛使用的中文排版软件主要包括北大方正的书版系列软件，飞腾系列软件，蒙泰桌面出版系统，Adobe 公司的 PageMaker，FrameMaker，以及 QuarkPress 公司的 PassPort 等。除此而外，由 D. E. Knuth 编写的 TEX 和由 L. Lamport 编写的 LATEX 也是学术界广泛的应用排版软件。

微软公司的 Word 是目前国内最为普及的字处理软件之一，也是大多数学校规定的学位论文编辑排版工具。不容否认，Word 在简单文书（例如：通知、简报等）编辑排版方面具有方便快捷的优势，而且其对多人协同编辑的支持也给文字修订工作带来了极佳的用户体验。但是从实际使用的情况看，尽管 Word 已经经历了第 12 个版本的改进，但是其对于处理大型文书文稿（例如：书籍、学位论文等）的能力仍然有待进一步完善和提高。由于 Word 版本不兼容造成的来回反复，也是使用 Word 编辑文字稿件的烦事之一。另外，由于 Word 对数学公式编辑的支持一直延续其“对象链接与嵌入”（Object Linking and Embedding，OLE）的设计理念，这也使得每位使用 Word 排版过理工类的文字资料的人都有一段或多段刻骨铭心的痛苦经历，往往花在调整格式这种 dirty work 上的时间和花在编写文章内容上的时间差不多或着甚至更多。

北大方正的书版系列软件是专业中文出版领域的权威，国内几乎所有的大型出版社、报社、政府机关几乎都使用书版系列软件对其出版的书籍、报纸和公文进行编辑排版。但是，书版软件作为方正电子出版流程中的一个主要组成部分，主要定位于印前排版环节，面向专业排版工作人员。因此，学习和使用使用书版软件需要花费较长的时间来熟悉复杂的排版命令，发排后需要使用专用的 RIP 软件或者方正的专用打印机才能输出样张等。

美国 Stanford 大学的荣誉退休教授 D. E. Knuth 在 197x 年独自一人开发了 TEX 排版系统，随后，L. Lamport 为 TEX 编写了一系列的宏包使得 TEX 的使用更加方便，这些宏包被称为 LATEX。自从 TEX/LATEX 问世以来它们就受到了学术界的青睐，目前几乎所有的国外出版社都接受或指定使用 TEX/LATEX 对稿件进行排版编辑。19xx 年，中国科学院的张林波研究员开发了 CCT 使得 LATEX 可以用于中文文稿的处理。德国的 W. Lemberg，编写了 CJK 宏包为 LATEX 提供了中日韩三国语言的解决方案。使用 TEX/LATEX 排版学术论文的最大优势在于，它让作者可以不用为排版输出的具体格

式操心，而全心投入文章、书稿内容的编写上，最大程度的降低作者从事排版 dirty work 的工作量。

目前，我国的清华大学、哈尔滨工业大学、西安电子科技大学、西安交通大学等都已经纷纷制作了本校学位论文的 LaTeX 模板，并接受使用 LaTeX 排版的学位论文。

1.2 本说明的主要内容

本说明全面介绍了如何使用 BUPTGraduateThesis 来排版符合 [2] 规定的北京邮电大学学位论文。全文内容安排如下：

1. 第二章介绍……
2. ……

参考文献

[1] 王惠敏. 大数据背景下电子商务的价值创造与模式创新 [J]. 2015(7): 76–77.

[2] 北京邮电大学研究生院培养与学位办公室. 关于研究生学位论文格式的统一要求 [EB/OL]. http://www.bupt.edu.cn/, 2014–11.

[3] 王旭. latex-bupt: LaTeXstyle for BUPT thesis[EB/OL]. http://code.google.com/p/latex-bupt/, 2009–01–14.

[4] 北京邮电大学研究生院培养与学位办公室. 关于研究生学位论文格式的统一要求 [EB/OL]. http://www.bupt.edu.cn/, 2004.

[5] 张煜. 北京邮电大学研究生学位论文 LaTeX 文档类 [EB/OL]. http://code.google.com/p/buptthesis/.

第二章　功能测试

脚注使用带圈数字的表示方法，此处为示例 1① 和示例 2②。

缩略语的功能非常强大，例如首次出现无线理论与技术实验室（Wireless Theories and Technologies Lab，WT&T） 和非首次出现 WT&T 时将显示不同的内容。

参考文献可以使用[1] 和 [2] 的表示方法。

2.1　三国演义

《三国演义》[3] 是中国第一部长篇章回体历史演义的小说，以描写战争为主，反映了蜀（汉）、魏、吴三个政治集团之间的政治和军事斗争，大致分为黄巾之乱、董卓之乱、群雄逐鹿、三国鼎立、三国归晋五大部分。

在广阔的背景下，上演了一幕幕波澜起伏、气势磅礴的战争场面，成功刻画了近五百个人物形象，其中曹操、刘备、孙权、诸葛亮、周瑜、关羽、张飞等人物形象脍炙人口，其中诸葛亮是作者心目中的“贤相”的化身，他具有“鞠躬尽瘁，死而后已”的高风亮节，具有近世济民再造太平盛世的雄心壮志，而且作者还赋予他呼风唤雨、神机妙算的奇异本领。曹操是一位奸雄，他生活的信条是“宁教我负天下人，休教天下人负我”，既有雄才大略，又残暴奸诈，是一个政治野心家阴谋家这与历史上的真曹操是不可混同的。关羽“威猛刚毅”、“义重如山”。但他的义气是以个人恩怨为前提的，并非国家民族之大义。刘备被作者塑造成为仁民爱物、视贤下士、知人善任的仁君典型。

2.1.1　长坂坡

京剧《长坂坡》[4] 是依据《三国演义》改编的京剧传统剧目。

故事叙述：刘备自烧屯新野之后，弃樊城，阻襄阳，一路率引军民，流离败走，穷促万分。关羽、诸葛亮，已先后遣往夏口，乞救于刘琦未返，刘备等往投江陵暂驻，中途经过当阳，驻扎景山之下。忽然曹操大兵，漫山遍野追至，夤夜厮杀，刘备众大败，及天明检点随从只余百余骑，刘备家眷及赵云、简雍、二糜等将，均不知下

① 测试脚注一
② 测试脚注二

落，其余百姓，亦均散失殆尽。此时赵云因于阿斗及甘、糜二夫人等失散，遂单骑冲突，四处找寻主眷，沓无下落。往回三数次，遇见简雍被创卧地，始略知失踪处所。赵云先救出简雍，令回，再往军中及百姓中搜访，先救甘夫人于难民队，同时又救糜竺，亲自护送至长坂坡，令糜竺保甘嫂先行，折身再回，觅糜嫂及阿斗。途中刺落夏侯恩，收获青釭宝剑，七次冲入重围，方得百姓指引，得见糜夫人抱阿斗坐于坍墙枯井之旁啼哭。夫人身受数创，不能行走。赵云叩见，极力请夫人上马，欲保护而出。夫人深知大义，惟以阿斗为托，己则以愿死报主，免累赵云，赵云再三安慰催行，力任无妨，夫人再三不可，亦促赵云速行。继见赵云坚待不去，恐且迟延遇寇，乃跳身入井，以速赵云之行。赵云大惊，尚踌躇设法营救，则曹军人马已至，不得已推墙掩井，解甲藏阿斗于胸前，忽忽上马，厮杀夺围欲出。此时曹操大兵云集，群矢于赵云一身，赵云在核心，东斩西杀，虽不败辱，而屡濒于厄。幸曹操爱勇将，赖徐庶乘间说曹操，以生擒勿伤，传令全军，始得完肤而返。

测试所有参考文献类型[5–13]。

参考文献

[1] 北京邮电大学研究生院培养与学位办公室. 关于研究生学位论文格式的统一要求 [EB/OL]. http://www.bupt.edu.cn/, 2014–11.

[2] 北京邮电大学研究生院培养与学位办公室. 关于研究生学位论文格式的统一要求 [EB/OL]. http://www.bupt.edu.cn/, 2004.

[3] 罗贯中. 山西太原: 元末明初.

[4] 赵云, 曹操, 刘备, 等. 长坂坡 [EB/OL]. http://baike.baidu.com/subview/428389/5476054.htm.

[5] Lippman S B, Lajoie J. C++ Primer 中文版 [M]. 王刚, 杨巨烽, 译. 第 5 版. 中国: 电子工业出版社, 2013: 1–838.

[6] Dahlman E, Gudmundson B, Nilsson M, et al. UMTS/IMT-2000 Based on Wideband CDMA[J]. IEEE Communications Magazine, 1998, 36(9): 70–80; year, volume, number and pages information for paper published in multi–series journals.

[7] Proceedings of IEEE Global Communications Conference (GLOBECOM'2008)[C]. New Orlean, USA: IEEE, 2008.

[8] Jindal N, Andrews J G, Weber S. Rethinking MIMO for Wireless Networks: Linear Throughput Increases with Multiple Receive Antennas[A]. // Proceedings of IEEE International Conference on Communications (ICC'2009)[C]. Dresden, Germany: IEEE, 2009: 1–6.

[9] Prasad N, Khojastepour M A, Jiang M, et al. MU-MIMO: Demodulation at the Mobile Station[R]. IEEE 802.16 Broadband Wireless Access Working Group, 2009: 1–11.

[10] TS 36.211 V10.5.0, Physical Channels and Modulation[S]. Valbonne, France: 3GPP, 2012-6.

[11] Paulraj A J, Heath R W Jr, Sebastian P K, et al. Spatial Multiplexing in a Cellular Network[P]. USA: 6067290, 2000–5–23.

[12] 吴刚. 立陶宛进入欧元时代 [N]. 人民日报, 2015–1–2.

[13] 百度百科. 香农公式 [EB/OL]. http://baike.baidu.com/view/747964.htm, 2013–10–28.

附录 A 不定型（0/0）极限的计算

定理 A.1 （L'Hospital 法则 ） 若

3. 当 $x \to a$ 时，函数 $f(x)$ 和 $g(x)$ 都趋于零；
4. 在点 a 某去心邻域内，$f'(x)$ 和 $g'(x)$ 都存在，且 $g'(x) \neq 0$；
5. $\lim\limits_{x\to a} \dfrac{f'(x)}{g'(x)}$ 存在（或为无穷大），

那么

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \lim_{x\to a} \frac{f'(x)}{g'(x)}. \tag{A-1}$$

证明： 以下只证明两函数 $f(x)$ 和 $g(x)$ 在 $x = a$ 为光滑函数的情形。由于 $f(a) = g(a) = 0$，原极限可以重写为

$$\lim_{x\to a} \frac{f(x) - f(a)}{g(x) - g(a)}.$$

对分子分母同时除以 $(x - a)$，得到

$$\lim_{x\to a} \frac{\dfrac{f(x) - f(a)}{x - a}}{\dfrac{g(x) - g(a)}{x - a}} = \frac{\lim\limits_{x\to a} \dfrac{f(x) - f(a)}{x - a}}{\lim\limits_{x\to a} \dfrac{g(x) - g(a)}{x - a}}.$$

分子分母各得一差商极限，即函数 $f(x)$ 和 $g(x)$ 分别在 $x = a$ 处的导数

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \frac{f'(a)}{g'(a)}.$$

由光滑函数的导函数必为一光滑函数，故 (A-1) 得证。 □

附录 B 缩略语表

BUPT	Beijing University of Posts and Telecommunications，北京邮电大学
WT&T	Wireless Theories and Technologies Lab，无线理论与技术实验室
WTI	Wireless Technology Innovation Institute，无线新技术研究所

致　谢

感谢 Donald Ervin Knuth.

攻读学位期间发表的学术论文目录

期刊论文

[1] **Zhang San** , Newton I, Hawking S W, et al. An extended brief history of time[J]. Journal of Galaxy, 2079, 1234(4): 567–890. （SCI 收录，检索号：786FZ）.

会议论文

[2] McClane J, McClane L, Gennero H, et al. Transcript in Die hard[A]. // Proc. HDDD 100th Super Technology Conference (STC 2046)[C]. Eta Cygni, Cygnus: 2046: 123–456. （EI 源刊）.

专利

[3] 张三 , 李四. 一种进行时空旅行的装置 [P]. 中国: 1234567, 2046–01–09.