* 大数据技术与软件前沿报告

摘要

随着科学技术的进步，计算机技术不断提高，互联网产生了大量的数据信息，人类已经进入了“大数据”时代，大数据是指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。现如今，大数据已经渗透到我们生活的方方面面了。本次报告将聚焦于大数据技术与软件的交互融合，旨在了解大数据时代软件的发展，以及与大数据技术结合的软件将会为国家，企业，用户带来什么。

英文摘要

With the progress of science and technology, computer technology, the Internet creates a large amount of data , mankind has entered the era of "big data", big data is not in a certain time range with conventional software tools to capture, manage, and deal with data collection. Big data is a massive, high-growth and diversified information asset, and we need new tools to deal with the big data that can help us to improve the ability of decision-making, insight and discovery, and the process optimization. Nowadays, big data has penetrated into every aspect of our lives. This report will focus on the interaction and integration of big data technology and software, aiming to understand the development of software in the era of big data, and what software combined with big data technology will bring to the country, enterprises and users.

简介

我国目前正处于大数据时代,随着科学技术的不断进步,大数据的应用也逐渐渗透到人们的日常生活中;大数据的融入推动传统软件开发发生变革。本次报告将聚焦于大数据技术与软件的交互融合，报告简要的讲述了大数据技术的起源、发展、现状等基本情况，大数据的主要发展方向，大数据背景下软件技术在社会中的具体应用研究。此外我们还概述了大数据技术在中国的发展，以及具体举例了大数据技术在视频app（腾讯视频）中的应用，最后我们概述了大数据技术的应用开发框架。

1. **介绍**

随着科学技术的进步,计算机技术不断提高,互联网已经产生了大量的数据信息,人类已经进入了“大数据”的时代,在这个时代软件产业发展迅猛,软件产品已经渗透到社会各领域的各个行业。并且由于软件系统产生的海量数据，信息系统中对数据的关注也正在迅速增加。例如，2015年，Youtube拥有超过10亿的用户，他们每天观看40亿个视频，每分钟上传300小时的视频。在软件应用使用产生的数据飓风中，软件工程师需要了解什么是大数据，如何开发大数据软件，大数据软件的需求是什么。

大数据是以数据容量为计量的一种全新的数据表达方式，容量大是其主要的发展特点，从目前对大数据的研究来看，大数据时代的“3V 理论”受到了大部分学者的认同，即：大数据应该具有“Volume”、“Velocity”以及“Variety”都较大的基本特征。这些基本特征是影响大数据在实际应用中的一个关键因素，从目前的信息发展来看，大数据的存在对计算机软件技术的发展应用提供了一定的数据处理支撑以及内容优化方式，随着软件技术数据规模的增大，计算机数据处理技术的要求提升是大数据时代的重要表现特征。现阶段，大数据相关的重要技术包括分类分析、机器学习、遗传算法等[1]。

今天，信息时代发展中对大数据的应用十分广泛，大数据分析已经成为每个行业的重要组成部分。大数据分析为重要的商业决策提供了洞见。谷歌、Facebook、IBM、Microsoft、Amazon、Oracle、Netflix等大型企业巨头都在投资大数据分析以获取商业优势，并创造了许多成功故事。正如人们所说，大数据是新兴产业的石油，因此，提炼石油以从中发现价值，而不仅仅是收集石油，同样重要。许多大小企业都面临着挑战和问题，需要开发能够从大数据中实时获益的软件。软件技术是以计算机信息技术为支撑经过延伸和发展逐渐应用于现代化社会的重要技术，当今社会正在处于经济结构转型的关键时期，与传统的手工生产模式相比，软件技术能够与生产模式相结合提高经济发展效率，为社会发展创造更多的经济效益。在生产领域的发展阶段，以提升核心竞争力为发展目的的生产运营工作的有效开展需要不同程度的将生产过程与软件化技术相结合，大数据时代的背景发展为软件技术在生活中的广泛应用提供了一定的数据基础，同时两者的有效结合发展能够为经济社会发展下各领域经济效益的有效提升创造了更多的条件。

1. **大数据的主要发展方向**

大数据是指海量的结构化和非结构化数据进行设计，通过水平可伸缩的软件处理不同的表示格式。从广义上讲，“大数据”被认为是一种社会经济现象，与分析大量数据的技术能力的出现有关，在一些问题领域和由此产生的变革效应。[2]定义大数据的特征有“三个V”：

Volume-作为物理体积的值；

Velocity-从增长率和对高速处理和获得结果的需要两方面考虑；

Variety-作为同时处理不同类型、结构化和半结构化数据的可能性。

大数据领域的主要技术有：

NoSQL技术：表示实现数据库存储的方法的数量。它们与传统关系数据库中使用的模型有很大的不同，可以通过SQL语言访问数据；

MapReduce技术：是一种分布式计算模型，用于在计算机集群中进行非常大的、高达几千兆字节的数据集的并行计算；

Hadoop技术：一套实用工具、库和框架，用于开发和执行在成百上千个节点的集群上运行的分布式程序；

R：一种统计数据处理和图形的编程语言，以及开源的自由软件计算系统；

商业智能(BI)技术：为组织提供计算机方法和工具的设计，这些方法和工具可将事务性商业信息转换为适合商业分析的人类可读形式，以及用于处理此类信息的大规模工作的手段。

1. **大数据背景下软件技术在社会中的具体应用研究**
   1. 商业化应用[1]

软件技术在商业领域的发展应用以及以软件技术为主要商业化发展方向的行业领域的发展都在飞速发展，可以看出，软件技术的应用于发展对于促进社会经济的有效转型以及经济结构的合理转换起到了积极的促进作用，一方面，软件技术的有效应用能够促进商业领域的高效管理与信息资源的有效整合，提升商业化竞争力，另一方面，软件技术还为生产行业的技术化生产提供了有效的生产途径，强化了生产过程的便捷性和有效性，促进生产效率的提升。以软件技术为商业化的发展方向更是从根本上促进了计算机软件的结构化创新，从软件的设计到软件的应用都不同程度的得到了资源的整合发展，提升了不同企业在资源化竞争中的有效竞争力，在促进信息的归纳整合与分析的同时，还增强了计算机软件技术的不断创新发展与高效利用。例如，以大数据发展背景下以软件技术为基础的网络商铺的应用，以信息安全技术为服务支撑的网络公司的发展以及以云储存技术为发展方向的互联网软件的广泛推广与应用，不仅增强了经济发展下的社会效益，还能够促进人们对资源的便捷利用与有效整合，提高了对数据的利用效率，发展了经济结构，促进了效益的提升。

* 1. 教学化应用[1]

软件技术在教育方向的发展应用是重要的着力点，教育是人才培养的基础和关键途径，软件技术的应用是提升教学效率，增强教学途径与方式的重要发展方向，再教育事业的发展中起着重要的推动作用，现阶段信息数字化的教学模式在教学过程中的广泛应用，能够有效避免传统教学方式下空间和时间的局限性，增强教育过程中学生的感知力与知识的整合过程，同时还能够有效提升教学对象的学习效率和兴趣程度。软件技术的应用不仅能够帮助教学主体对教学资源进行结构整合与资源利用，还能够实现教学内容由抽象到具体的合理转变，帮助学生对课堂教学内容进行合理的理解和有效的消化，从而提高教育教学中的效率。大数据背景下软件技术的应用能够对学生的学习数据进行有效的整合，通过分析发现不同学生在学习阶段的闪光点与学习弱点，进而实现教师在教学阶段的针对化辅导，在促进教学手段提升的同时还增强了教学效果，促进了教育内容向数据资源的有效转化，实现了学习资源的多向高效利用。

* 1. 通信化应用[1]

通信领域的发展离不开计算机软件技术的支持，借助于IBMSPSS软件的预测和分析，能够充分地提升对各种商业信息的分析准确性，进而减少没必要的信息损失。在通信化领域利用软件技术对顾客的行为进行精准化的数据分析与合理化的预测，能够促进企业在发展过程中对问题的有效分析和反馈，从某种角度来看，通信行业与软件行业是相辅相成的，通信行业的数据量相对庞大，将软件技术与其进行结合能够实现对消费者客户信息的有效分析，不仅能够对消费喜好与购买能力进行分析，还能够对未来的购买方向进行精准预测，通过这些数据信息将客户进行合理的定位，能够促进企业发展中经济效益的最大化发展。例如，在通信行业以客户在流量的使用情况以及消费情况为数据进行分析，能够精准得出不同消费人群在套餐方面的使用情况，进而通信企业可以根据消费情况推出不同类型的消费套餐供客户选择，在提高服务精准化的同时还增强了企业的经济效益，促进了经济社会的高效转型。大数据背景下软件技术的应用在增强企业的发展效益与经济效益的同时，还提高了服务质量的精准化程度，促进了社会的和谐稳定发展。综上所述，经济社会的发展与大数据背景下软件技术的发展是相辅相成的关系，经济社会的飞速发展促进了软件技术的不断优化提升，软件技术在企业行业的发展应用促进了经济结构有效转型发展，但同时，应当充分认识当今社会发展下计算机软件技术面临的挑战和难题，并在未来的发展中不但优化提升，促进经济社会转型与大数据背景下软件技术的高效协同发展。

1. **大数据技术在中国的发展和成长**

习近平总书记在十九届中共中央政治局第二次集体学习时的重要讲话中指出：“大数据是信息发展的新阶段”，并做出了“推动大数据技术产业创新发展、构建以数据为关键要素的数字经济、运用大数据提升国家治理现代化水平、运用大数据促进保障和改善民生、切实保障国家数据安全”的战略部署，为我国构筑大数据时代国家竞争新优势指明了方向。可以说，大数据技术的发展在中国是得到了全方面的支持的。

大数据于2012、2013年达到其宣传高潮，2014年概念体系逐渐成形，对其认知亦趋于理性。大数据相关技术、产品、应用和标准不断发展，逐渐形成了包括数据资源与API、开源平台与工具、数据基础设施、数据分析、数据应用等板块构成的大数据生态系统，并持续发展和不断完善，其发展热点呈现了从技术向应用、再向治理的逐渐迁移[9]。经过多年来的发展和沉淀，人们对大数据以及形成基本共识：大数据现象源于互联网及其延伸所带来的无处不在的信息技术应用以及信息技术的不断低成本化。大数据泛指无法在可容忍的时间内用传统信息技术和软硬件工具对其进行获取、管理和处理的巨量数据集合，具有海量性、多样性、时效性及可变性等特征，需要可伸缩的计算体系结构以支持其存储、处理和分析。

在我国，大数据的现状有以下三方面。一是我国已有众多成功的大数据应用，但就其效果和深度而言，当前大数据应用尚处于初级阶段，根据大数据分析预测未来，指导实践的深层次应用将成为发展重点。二是大数据治理体系远未形成，特别是隐私保护、数据安全与数据共享利用效率之间尚存在明显矛盾，成为制约大数据发展的 重要短板，各界已经意识到构建大数据治理体系的重要意义，相关的研究与实践将持续加强。三是数据规模高速增长，现有技术体系难以满足大数据应用的需求，大数据理论与技术远未成熟，未来信息技术体系将需要颠覆式创新和变革。

针对我国大数据挖掘与分析能力弱、大数据算法应用和综合能力不高等问题，急需开展关键技术的攻关研发与突破工作。为践行国家大数据战略，北京大学牵头，联合中国科学院数学与系统科学研究院、中山大学、北京奇虎科技有限公司、北京嘀嘀无限科技发展有限公司、中国信息安全研究院有限公司等多家单位，共同申请建设大数据分析与应用技术国家工程实验室，并与2017年初获得了国家发展和改革委员会的正式批复[10]。其目标是围绕数据科学理论体系、大数据计算系统与分析理论等重大基础研究进行前瞻布局，开展数据科学研究，引导和鼓励在大数据分析的理论、方法及关键应用技术等方面展开探索；加强数据清洗，大数据分析挖掘、大数据可视化等领域关键技术攻关；围绕多数异构数据分析，针对视频、文本等多种数据类型，研究大规模机器学习的基础和创新方法，依托合作支撑单位开展面向大数据分析的全流程框架和工业标准的设定、大规模数据分析系统平台的搭建等。

1. **大数据技术如何应用到软件产品以及软件产品范例**

大数据技术已经在多个app有所体现，比如腾讯、爱奇艺、bilibili等视频软件，比如近日头条等新闻app，比如网易云音乐、qq音乐等音乐软件，比如淘宝、京东等购物app等等。

在视频APP上的大数据分析可以分成两部分，一个在用户侧，主要分析用户行为；另一个在内容侧，分为两方面，一是对内容本身的分析，另一个是用户在内容上发生的行为的分析。本节中将以腾讯视频为例子，讲解大数据技术在视频软件中的应用。

首先以一个未登录的游客身份打开腾讯视频，它的首页如图1所示。

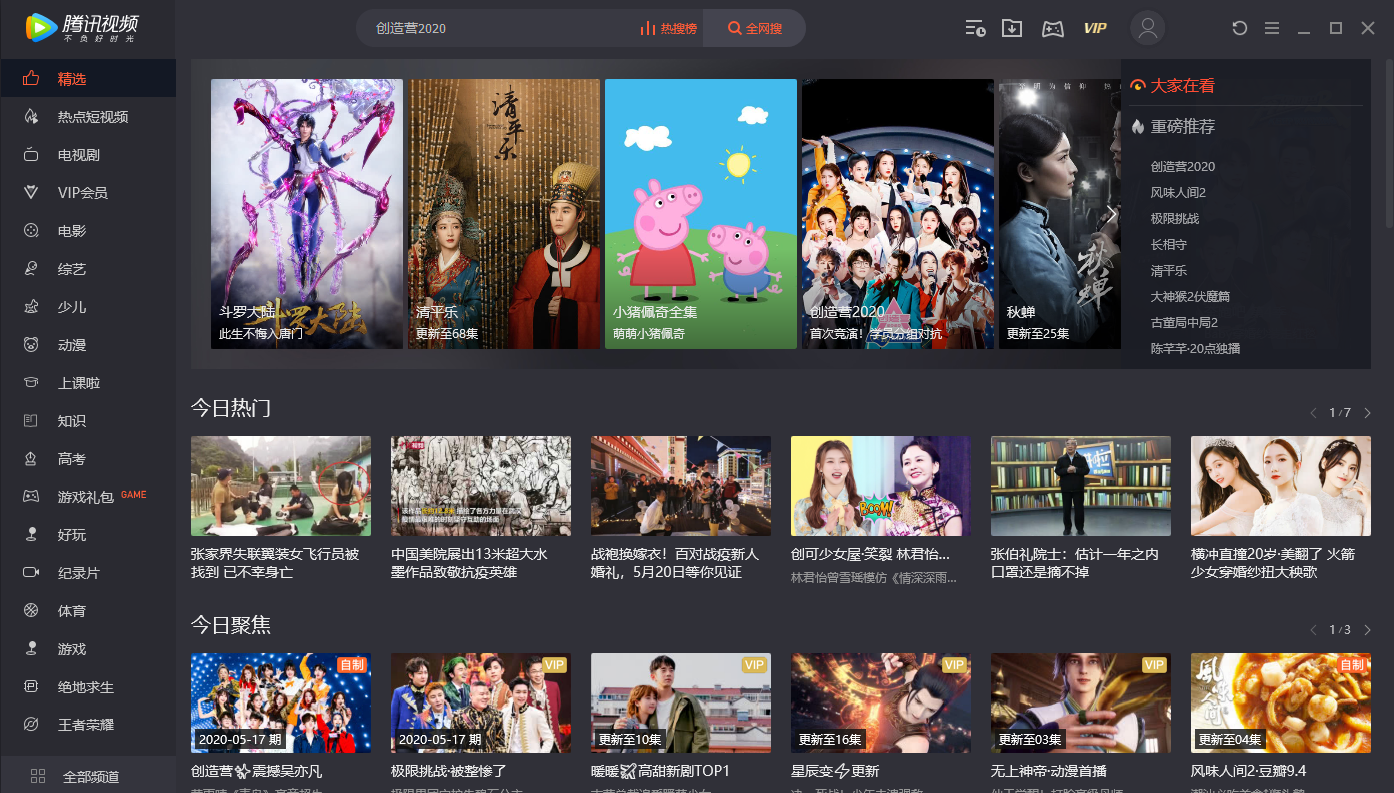


图1 用户未登录时的视频软件主页

用户登陆之后的界面如图2所示

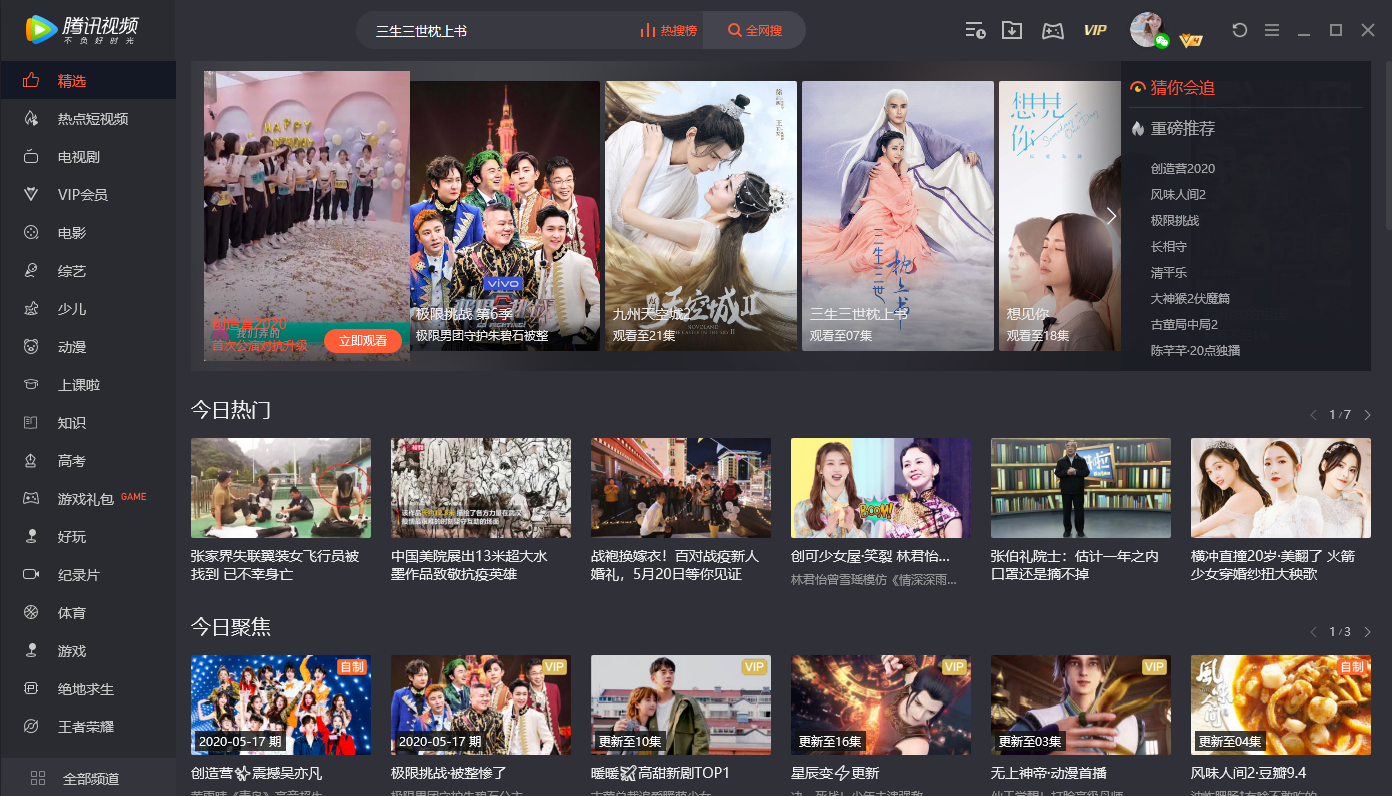


图2 用户登录后的视频软件主页

将图1和图2作对比。可以发现，当用户没有登录时，首页会出现一个模块叫“大家在看”，而登录以后相同位置的模块是“猜你会追”，“大家在看”中的推荐的视频的依据是腾讯视频所有用户正在观看的几个最热的视频，“猜你会追”中出现的视频其实就是该用户以往点击过或者看过的视频。然后点击“猜你想追”，会跳到图3所示界面。

腾讯视频会根据用户最近看过的视频或者用户搜索的关键词，将相关视频放在“猜你在追”这个页面，然后合理猜测，用户是因为喜欢这个视频里的某位明星，或者是这个视频的类型比较吸引该用户，然后它会列出这个明星相关的影视剧，也会列出同类型的视频，如图3、图4所示。

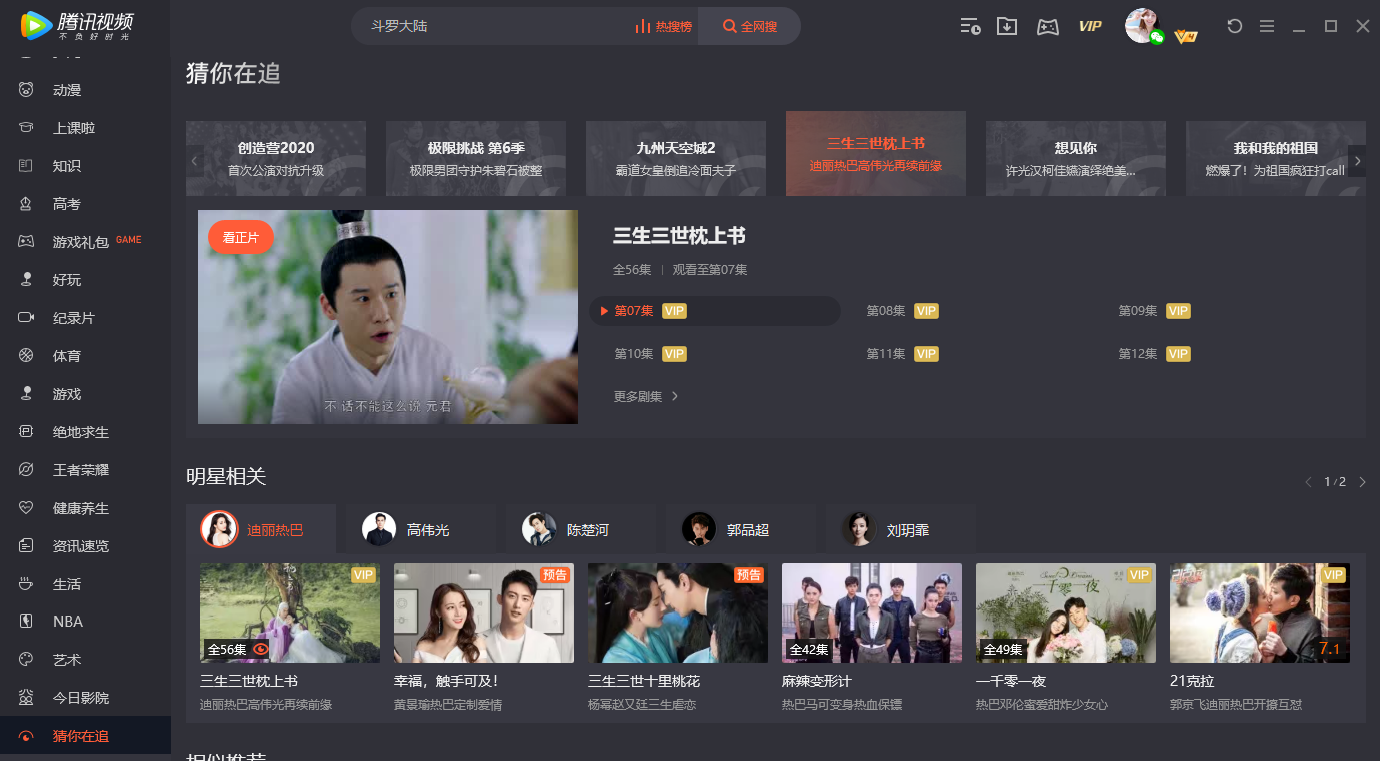


图3 “猜你想看”界面



图4 “相似推荐”界面

如果用户这次的浏览点的是这个明星相关的视频，并且接着观看了这个明星相关的一系列的视频，那么腾讯视频会在用户经常观看的这个明星上增加权重。如果用户这次的浏览点击是同类型的视频，那么腾讯视频就会在用户经常观看的视频类型上增加权重。若一开始的某个信息的权重一直降低，那么最后就会抛弃这个信息。因此，当一个用户在腾讯视频上花费的时间越多，腾讯视频就能收集到更多这个用户的数据多，就能更加了解用户的喜好，之后可以更好地推荐用户喜欢的视频，增加用户粘度。当然腾讯视频也不是说100%推荐用户喜欢的视频，它会推荐其中80%，然后用剩下的20%去挖掘用户可能喜欢的之前没有关注过的别的类型的视频，否则当一个用户在这个视频app上观看完所有他喜欢的那个类型的视频以后，会面临着无剧可看的尴尬场面，可能也会马上就放弃这个视频app。

当然，视频软件中的大数据内容分析不仅仅包括上面说的，大数据分析涵盖多个角度，比如说，用户观看的时间段偏好、用户投入度、观看时长分布、情节点分析、衍生流量的贡献、弹幕评论关键词、播放完成率等等[11]。

大数据在内容分析中包括评估、制作、运营、评价等四方面[12]。现假设腾讯视频想要制作一部剧，于是它会有一个IP评估，根据这部剧以往的流量来看，这个IP适不适合拍成电视剧。其次还要选演员，选择档期合适，片酬合适的演员。以往这些评估需要大量的人力去做，但在大数据时代，这个评估就变得更有依据性，结果也更不容易出错。还可以很合理的分配演员片酬、制作经费、宣传费用等制作费用的占比。

1. **大数据技术的应用开发框架概述**

伴随着移动技术的迅速发展，各国互联网企业开始思考要如何才能够有效地控制移动应用的成本投入，构建出一个完整的、健全的生态链，在版本更新换代的背景下，深入发掘用户的需求，提高用户的使用体验。按照当前国内的情况来看，在2018年初我国移动应用开发框架以及开发工具迅速发展，出现了大量功能全面的开发框架以及开发工具，为了适应市场的发展需求，各大厂商把成本控制和满足客户需求作为企业生产与发展的战略目标，应用开发框架化受到了社会各界的关注[13]。

主流框架[14][15]

1. Bootstrap移动应用开发框架。Bootstrap是以html、css、java为主体的框架，它是在jquery基础上进行优化和调整后所形成的前端开源框架。优点是兼容性好，缺点是对于移动应用动态加载的效果差。
2. Amaze UI移动应用开发框架。该框架是以CSS3技术为核心，利用js、web组件为辅的框架设计结构，该框架的优点是自由度高，支持用户自定义字体与内容，缺点是缺少动态加载技术。
3. JQuerymobile移动应用开发框架。该框架是以JQuery与用户界面结合的基础上开发出的框架，该框架具有自定义主题、支持高端设备使用的优点，但该框架在显示及分辨率方面存在适配问题，且不支持动态加载技术。

参考文献

[1] 徐立宇.大数据背景下软件技术的发展[J].电子技术与软件工程,2019,12:162-163.

[2] TAN Chaodong, ZHANG Hengru, MA Yongzhong, et al. Research and Application of Big Data Mining System for Oil-gas Production[J], Peak Data Science. 2016. 5(1):49-51.

[3] Alexey V Bataev. Analysis of the Application of Big Data Technologies in the Financial Sphere[J]. Saint-Petersburg,2018,12:568-572.

[4] W. Shang, B. Adams, and A. E. Hassan, "An experience report on scaling tools for mining software repositories using mapreduce," in Proceedings of the IEEE/ACM international conference on Automated software engineering, 2010, pp. 275-284.

[5] Gorton, I. Bener, A. B. and Mockus, A. (2016). Software Engineering for Big Data Systems. IEEE SOFTWARE. pp: 32-35.

[6] T.H.A.S. Siriweera, I. Paik, B. T.G.S. Kumara, K.R.C. Koswatta, “Intelligent big data analysis architecture based on automatic service composition,” in Big Data(Big Data Congress), 2015 IEEE Int. Congress on, pp.276-280. IEEE, 2015.

[7] A. A. Munshi, Y. Abdel-Rady I. Mohamed, “Data Lake Lambda Architecture for Smart Grids Big Data Analytics,” IEEE Access 6 (2018): 40463-40471.

[8] D. Bollier and C. M. Firestone, The promise and peril of big data. Washington, DC: Aspen Institute, Communications and Society Program, 2010.

[9] 胡声艳. 试谈大数据时代背景下计算机软件技术的应用[J]. 电脑编程技巧与维护, 2018, (11):104-105，116.

[10] 张平文, 鄂维南, 袁晓如等. 大数据分析与应用技术创新平台[J]. 大数据, 2018, (4):86-93.

[11] 陆治. 浅谈大数据在视频网站的应用[EB/OL]. 2018, <https://max.book118.com/html/2018/0714/8101007066001115.shtm>

[12] 张文. 大数据分析对中国影视业运营的意义[EB/OL]. 2018, https://ishare.iask.sina.com.cn/f/135KSiRk6IM.html

[13] 尤丹妹, 魏慧琴. 基于移动平台的学习资源设计与开发[J]. 微型机与应用, 2015, 第34卷(7):83-85，89.

[14] 段锦. 基于大数据背景下移动平台的微应用开发框架技术的研究[J]. 营销界, 2019(38).

[15] 查英华,朱其慎.基于个性化推荐的移动学习系统构建[J].新课程研究（中旬-单）,2015,(5):7-9.