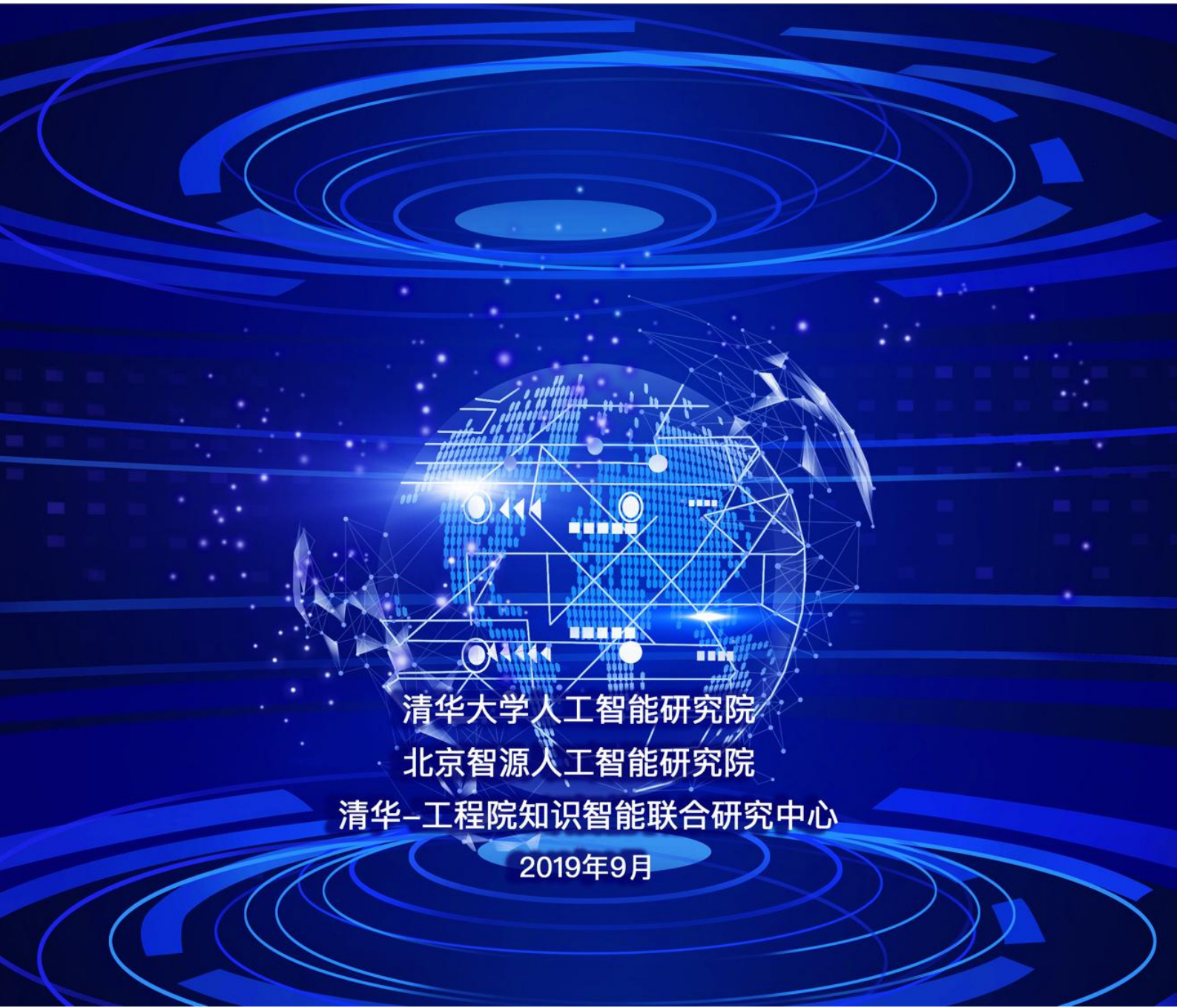
THUAITR

人工智能之信息检索与推荐

Research Report of Information Retrieval and Recommendation

2019年 第5期



清华大学人工智能研究院
北京智源人工智能研究院
清华–工程院知识智能联合研究中心

2019年9月

目录

CONTENTS

图目录

表目录

1 概述篇

我们生活在一个信息时代，并正朝着数字化时代迈进。信息社会化、社会信息化、信息生产与消费促进了信息产业和信息技术的飞速发展，尤其是互联网的发展。然而互联网规模和信息资源的迅猛发展带来了信息过载的问题，一方面人们可以获取海量信息，另一方面信息获取的成本却提高了，人们不仅需要查询信息，还要剔除自己不需要的信息。因此，信息检索与推荐技术应运而生。信息检索技术可以帮助用户快速查找所需信息，满足用户的主流需求，而推荐技术能够在分析预测用户需求的基础上推送用户们可能需要但又无法获取的有用信息，提供个性化服务。

信息检索系统与信息推荐技术的产生和发展有效地提高了用户们获取信息的效率，优化了信息服务系统。随着人们对信息化技术的依赖加强，信息检索与推荐将会朝着更加智能化、个性化、专业化的方向发展，成为人们筛选、浏览信息时的必备工具。

1.1 信息检索的概念与发展

1.1.1 信息检索的概念

信息检索（Information Retrieval，IR）是指信息的表示、存储、组织和访问。信息检索有广义和狭义之分。广义的信息检索，包括信息存储与检索，是指信息按一定的方式进行加工、整理、组织并存储起来，再根据信息用户的需要将相关信息准确的查找出来的过程。

狭义的信息检索仅指信息查询（Information Search），即用户根据需要，借助检索工具，提出查询要求，数据库匹配出与之有关的资料。

信息检索的主要环节包括信息内容分析与编码、组成有序的信息集合以及用户提问处理和检索输出。其中信息提问与信息集合的匹配、选择是整个环节中的重要部分。当用户向系统输入查询时，信息检索过程开始，接着用户查询与数据库信息进行匹配。返回的结果可能是匹配或不匹配查询，而且结果通常被排名。大多数信息检索系统对数据库中的每个对象与查询匹配的程度计算数值分数，并根据此值进行排名，然后向用户显示排名靠前的对象，信息检索框架如图 1 所示[1]。

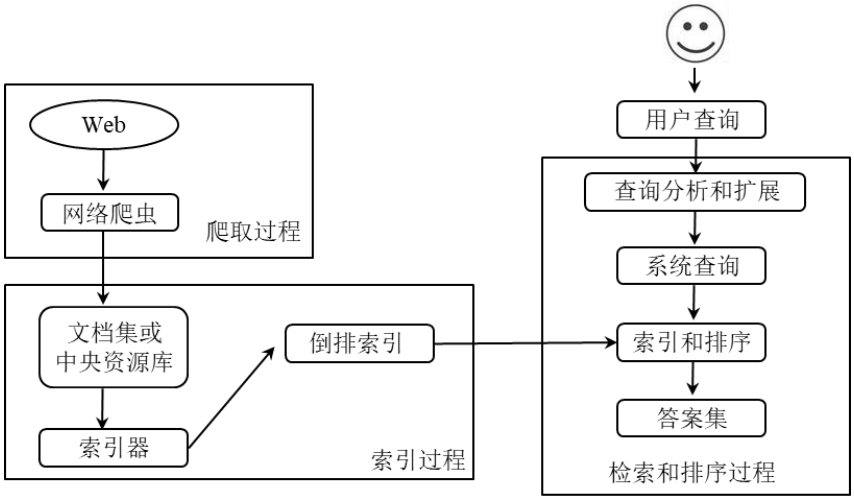


图 1 信息检索系统构架

1.1.2 信息检索的发展历程

信息检索的目的是获取所需信息，而这要基于比较完善的检索技术[2]，用户需求的变化和信息技术的进步对信息检索的发展有着重要的影响。根据技术的演化，我们将信息检索发展历程分为三个阶段：

（1）数字图书馆 / 文档电子化时代

1954 年，Vannevar Bush（范内瓦·布什，图 2）在“Atlantic Monthly”7 月号发表了一篇名为“As We May Think”的文章，这篇文章影响了几代的计算机科学家。文章提到：“未来人们能够实现对海量图书资源（1M）进行快速的访问”。概括出了信息检索在数字图书馆时代的特征，即对文档全文内容的快速检索。

范内瓦·布什在担任美国科学研究与发展办公室主任期间推进了美国军队研究机构与高校研究机构的合作，正是当时在这种合作关系中发挥最重要影响的三所大学（哈佛大学、麻省理工学院、加州大学伯克利分校）与后来成立的美国国防部高等研究计划署（ARPA）合作开发出了互联网的雏形：ARPANET。

图 2 范内瓦·布什（1890-1974）

1957 年，Luhn 在论文“A Statistical Approach to Mechanized Encoding and Searching of Literary Information”里提到“…a writer chooses that level of subject specificity and that combination of words which he feels will convey the most meaning.”这是一种以单词作为索引单元的文档检索方法。

20 世纪 60 年代，Gerard Salton 创造了信息检索系统“SMART”（Salton’s Magic Automatic Retrieval of Text），推进了信息检索相关研究的水平提升。SMART 系统并非搜索引擎，但它具备搜索引擎具有的文本索引、查询处理、结果排序等功能。

20 世纪 60 年代后期另外两个研究领域需要提及。第一个是 Julie Beth Lovins 于 1968 年在麻省理工学院开发的词干算法（Stemming Algorithm）；另一个研究涉及评估指标，例如 William Cooper 在 1968 年提出的“Cooper”，这个度量标准目前已在多个应用程序中大量使用。

在数字图书馆时代，信息检索技术主要应用于封闭数据集合、单机模式或专网内的主机-终点模式，在商业应用方面，则是提供软件/解决方案，专网内的查询服务。

（2）早期互联网时代

随着信息技术的爆炸式发展，信息检索的发展发生了质的飞跃。Tim Berners-Lee（蒂姆·伯纳斯·李，图 3）基于尚未被商用的互联网提出了万维网（Web）的原型建议。1991 年 8 月，蒂姆·伯纳斯·李在一台 NeXT 电脑上建立了第一个网站 http://nxoc01.cern.ch/。他一直坚持将公开和开放作为万维网的灵魂。

图 3 蒂姆·伯纳斯·李和他的 NeXT 电脑

1994 年第一届 WWW（International World Wide Web Conference）会议召开，借助 Hyper-text（超链接文本）、Links（链接）和 Connected Web（网络）的万维网能够把不同电脑上的文本、图像、声音等链接起来，使得“链接一切”成为了可能，信息检索由此进入了早期互联网时代，即以链接分析为代表的大规模 Web 搜索。

在这个时期，学术界和业界都发生了深刻变化。国际上开始细分不同的检索任务的评价方法和探讨大规模 Web 数据的评测标准。国内在 2003 年召开了第一届全国搜索引擎和网上信息挖掘学术研讨会；2004 年召开了第一届全国信息检索与内容安全学术会议；2006 年 11 月 21 日成立了信息检索与安全专委会。业界主要表现为第一代搜索引擎和第二代搜索引擎的出现，国外有 AltaVista、Excite、WebCrawler 和 Yahoo！国内有应用于国防和安全领域的“天罗”，和面向公众提供服务的天网。第二代搜索引擎的代表是 1998 年成立的 Google 和 2000 年 1 月创建的中文搜索引擎--百度，在百度之后，多家中文搜索引擎相继出现，例如，中搜、搜狗、搜搜和有道。

这个时期信息检索的应用形态的特征是开放的、大规模的、实时的、多媒体的。尤其巨型搜索引擎采集到的公开数据和用户访问日志等非公开数据深刻地影响着这一时期信息检索领域的创新模式。

（3）Web2.0 时代

在 Web2.0 时代，用户对 Web 有更深入的参与需求，这就对信息检索提出了更高的要求。信息搜索的发展开始更加关注用户需求，以实现内容与行为的精准 Web 搜索。

这个时期的信息检索实现了内容数据与社会各侧面的电子化数据（万维网、社交网、物联网、地理信息等）的全面融合；尤其是对社交网络数据的采集和大数据处理技术出现了社会化趋势。

1.2 信息推荐的概念与发展

1.2.1 信息推荐的概念

互联网规模和信息资源的迅猛增长带来了信息过载的问题，如何获取所需信息日益困难。以“信息推送”为服务模式的信息推荐系统，是当前解决信息过载问题的主要手段。信息推荐（Information Recommendation）是指系统向用户推荐用户可能感兴趣但又无法获取的有用信息，它的实现主要依靠推荐系统。

信息推荐的系统架构和运行方式吸收了信息检索系统设计中许多有价值的经验，例如文档处理与查询处理过程与传统信息检索系统的运行原理。总体而言，搜索引擎系统由数据抓取子系统、内容索引子系统、链接结构分析子系统和内容检索子系统四个组成部分构成[3]，如图 4 所示。

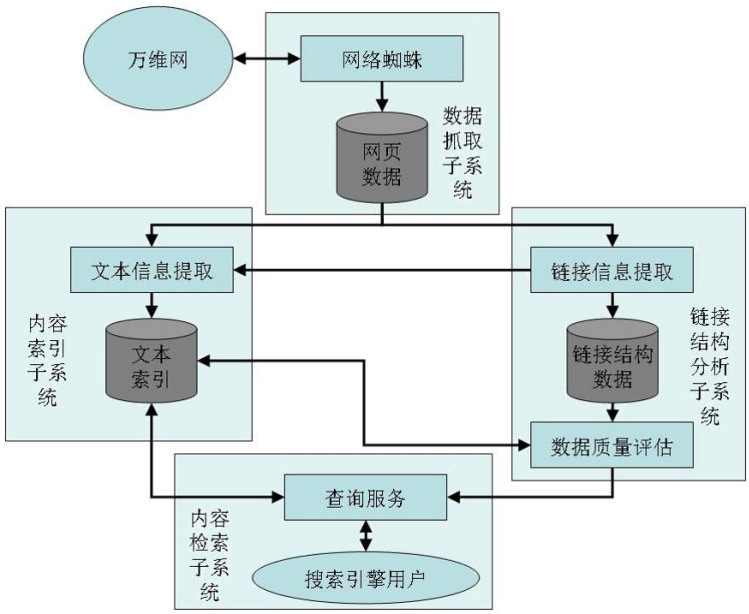


图 4 信息推荐系统架构示意图

1.2.2 信息推荐的发展历程

上个世纪最后二十年以来互联网的发展和普及为人们提供了一个全新的信息存储、加工、传递和使用的载体，网络信息也迅速成为了社会成员获取知识和信息的主要渠道之一。

一般认为推荐系统（Recommender System）的研究始于 1994 年明尼苏达大学，Group Lens 研究组推出的 Group Lens 系统，该工作不仅首次提出了协同过滤的思想，并且为推荐问题建立了一个形式化的模型，为随后几十年推荐系统的发展带来了巨大影响。

之后，推荐系统的相关技术得到了进一步发展和重视。1995 年 3 月，卡耐基·梅隆大学的 Robert Armstrong 等人在美国人工智能协会提出了个性化导航系统 Web Watcher；斯坦福大学的 Marko Balabanovic 等人在同一会议上推出了个性化推荐系统 LIRA；1997 年，AT&T 实验室提出了基于协作过滤的个性化推荐系统 PHOAKS 和 Referral Web；2000 年，NEC 研究院的 Kurt 等人为搜索引擎 CiteSeer 增加了个性化推荐功能；2003 年，Google 开创了 AdWords 盈利模式，通过用户搜索的关键词来提供相关的广告。2007 年开始，Google 为 AdWords 添加了个性化元素，不仅仅关注单词搜索的关键词，而且对用户一段时间内的推荐历史进行记录和分析，据此了解用户的喜好和需求，更为精确地呈现相关的广告内容；2009 年 7 月，国内首个推荐系统科研团队北京百分点信息科技有限公司成立，该团队专注于推荐引擎技术与解决方案，在其推荐引擎技术与数据平台汇集了国内外百余家知名电子商务网站与资讯类网站，并通过这些 B2C 网站每天为数以万计的消费者提供实时智能的商品推荐。

信息推荐系统的演变始终伴随着网络的发展，第一代信息推荐系统使用传统网站从以下三个来源收集信息：来自购买或使用过的产品的基础内容数据；用户记录中收集的人口统计数据；以及从用户的项目偏好中收集的基于记忆的数据。第二代推荐系统通过收集社交信息，例如朋友、关注着、跟随者等）。第三代推荐系统使用网上集成设备提供的信息。

信息推荐系统刚开始专注于通过过滤来提高推荐准确性，开发并优化了大多数基于存储器的方法和算法，在这个阶段，混合方法提高了建议的质量。在第二阶段，调整和开发了包括具有先前混合方法的社交信息的算法。作为一种人机交互系统，信息推荐系统已经广泛应用于社会生活的各个方面，因此系统地探讨信息推荐系统的发展历程具有重要意义。

1.3 信息检索和信息推荐的联系和区别

信息的检索与推荐都是用户获取信息的手段，无论是在互联网上，还是在线下的生活场景里，这两种方式都大量并存，两者之间的关系是互补的：搜索引擎需要用户主动提供准确的关键词来寻找信息，因此不能解决用户的很多其他需求，比如当用户无法找到准确描述自己需求的关键词时，搜索引擎就无能为力了。和搜索引擎一样，推荐系统也是一种帮助用户快速发现有用信息的工具。和搜索引擎不同的是，推荐系统不需要用户提供明确的需求，而是通过分析用户的历史行为给用户的兴趣建模，从而主动给用户推荐能够满足他们兴趣和需求的信息。因此，从某种意义上说，推荐系统和搜索引擎对于用户来说是两个互补的工具。搜索引擎满足了用户有明确目的时的主动查找需求，而推荐系统能够在用户没有明确目的的时候帮助他们发现感兴趣的新内容。在实际生活中也有很多运用。例如很多互联网产品不仅提供搜索功能，还会根据用户的喜好进行推荐，例如，对提供音乐、新闻或者电商服务的网站，必然要提供搜索功能，当用户想找某首歌或某样商品的时候，输入名字就能搜到；与此

同时，也同时要提供推荐功能，当用户只是想听好听的歌，或者打发时间看看新闻，但并不明确一定要听哪首的时候，给予足够好的推荐，提升用户体验。

同时，信息的检索与推荐也有着一定的区别，可以分为以下几个方面：

首先是主动与被动的不同。搜索是一个非常主动的行动，用户的需求也十分明确，在搜索引擎提供的结果里，用户也能通过浏览和点击来明确的判断是否满足了用户需求。然而，推荐系统接受信息是被动的，需求也都是模糊而不明确的。

其次是个性化程度的高低。搜索引擎虽然也可以有一定程度的个性化，但是整体上个性化运作的空间是比较小的，是当需求非常明确时，找到结果的好坏通常没有太多个性化的差异。但是推荐系统在个性化方面的运作空间要大很多，虽然推荐的种类有很多，但是个性化对于推荐系统是如此重要，以至于在很多时候大家干脆就把推荐系统称为“个性化推荐”甚至“智能推荐”了。

再次就是需求时间不同。在设计搜索排序算法里，需要想尽办法让最好的结果排在最前面，往往搜索引擎的前三条结果聚集了绝大多数的用户点击。简单来说，“好”的搜索算法是需要让用户获取信息的效率更高、停留时间更短。但是推荐恰恰相反，推荐算法和被推荐的内容往往是紧密结合在一起的，用户获取推荐结果的过程可以是持续的、长期的，衡量推荐系统是否足够好，往往要依据是否能让用户停留更多的时间，对用户兴趣的挖掘越深入，越“懂”用户，那么推荐的成功率越高，用户也越乐意留在产品里。

最后是评价方法不同。搜索引擎通常基于搜索引擎通常基于 Cranfield 评价体系，整体上是将优质结果尽可能排到搜索结果的最前面，让用户以最少的点击次数、最快的速度找到内容是评价的核心。而推荐系统的评价要宽泛很多，既可以用诸如 MAP（Mean Average Precision）的常见量化方法评价，也可以从业务角度进行侧面评价[4]。

2 技术篇

随着信息产生媒体和载体的多样化，网络环境中的信息种类越来越多，信息总量不断增长，内容复杂多样。如何快速的获取信息，准确的将信息推荐给用户，急需相应的理论和技术来支持，利用相应的理论方法和技术手段汇集、过滤、存储、推荐信息，方能满足用户信息查询和获取的需要，提高信息的利用效率。本章遴选部分信息检索与推挤的相关技术，从技术内容的角度对信息检索和推荐进行介绍。

2.1 信息检索部分前沿技术

2.1.1 集合论模型

2.1.1.1 布尔模型

布尔模型是基于集合论和布尔代数的一种简单检索模型，是早期搜索引擎所使用的检索模型[5]。它的特点是查找那些对于某个查询词返回为“真”的文档。在该模型中，一个查询词就是一个布尔表达式，包括关键词以及逻辑运算符。通过布尔表达式，可以表达用户希望文档所具有的特征，例如必须包含哪些关键词，不能包含哪些关键词等等。例如我们希望查找那些既含有“清华”又含有“大学”的网页，那么查询词可以写作“清华 AND 大学”。由于文档必须严格符合检索词的要求才能够被检索出来，因此布尔检索模型又被称为“完全匹配检索”（Exact-Match Retrieval）。

传统的布尔检索是将用户查询与文献进行逻辑的(而非数值的)比较而获得结果的检索。布尔检索模型的突出优点在于这种结构化的提问方式与用户的思维习惯相一致。同时，这种模型把复杂的检索过程简单化，能够将较复杂的情报提问按其概念组面的逻辑关系描述出来，从而变成可以由计算机执行的逻辑运算，变成机器根据事先确定的程序进行自动匹配的过程，这种运算上的简单易行是布尔检索系统的又一突出特征。此外，用布尔检索进行操作的某些系统允许用户通过给他使用的一个有结构的词典来缩小或扩大检索。所谓有结构的词典是指对任何一个给定的标引词都存储了与之相关的更一般的（上位）或更精确的（下位）关键词的词典。布尔检索很容易利用这些相关项来改进检索。

布尔检索在理论上存在的一些缺陷也是不容忽略的，具体包括下列几个方面。

（1）布尔逻辑式的构造不易全面准确反映用户的需求。

（2）匹配标准存在不合理的地方，严格的匹配可能导致检出的文档过多或过少，难以控制结果输出量的大小。

（3）对检索结果平等对待，不能按照用户定义的重要性排序输出。

（4）对用户的检索技能有较高的要求。

2.1.1.2 扩展布尔模型

布尔检索简单优雅，然而，由于它不支持索引项权重，因此它也不生成答案集的排序，故而输出的规模可能过大或者过小。由于这些问题，现代信息检索系统不再基于布尔模型。实际上，大部分的新系统其核心采用某种形式的向量检索。其原因是向量空间简单、快速，能产生更好的检索质量。另一种方法是用部分匹配和项权重的功能来扩展布尔模型。这种方法可以使人们可以把布尔查询表达式与向量。

考察一个合取布尔查询 $q=k_x \wedge k_y$。根据布尔模型，一篇仅包含 k_x 或者 k_y 其中之一的文档与另一篇不包含其中任何一个的文档都是不相干的。然而，这种决策通常与常识不符。鉴于此，Salton、Fox、Wu 在 1982 年引入了扩展布尔模型[6]。扩展布尔模型扩展了布尔代数，用代数距离来解释布尔操作符。在此意义上来讲，扩展布尔模型是用向量模型的特征来扩展布尔模型。

2.1.1.3 模糊集模型

用模糊集来表示模糊性与不确定性是有价值的，Ogawa、Morita 等人将其应用于信息检索领域[7]。文档的信息检索过程实际上涉及文档集的表示、用户查询的表示、相似性匹配及其排序三部分。首先，文中用户查询和文档集的表示均采用下列方式表示：$A=\{x_i/\mu A(x_i), x_i \in U\}$，对于文档集中的 x_i 为从检索词论域 U 中提取出来的能够代表整个文档意思的检索词集，$\mu A(x_i)$为提取出来的检索词属于该集合的隶属度，可以理解为每个检索词 x_i 属于该集的权重。对于用户查询中的 x_i 的解释同文档集中的 x_i，其中的$\mu A(x_i)$同样可以理解为权重，或者是该检索词的兴趣度。其次，基于上述给出的主导隶属度函数关系可知，只要查询中的检索词隶属度小于文档中的检索词隶属度，那么查询检索词集就包含于文档集，通过这一点就能找出包含某一查询检索词集的所有文档。这就是文档和查询的匹配。也就是说当给出了某查询检索词集，通过包含度定理计算其包含于文档的程度，根据这个包含度的大小来对检索出来的文档进行排序。

2.1.2 代数模型

2.1.2.1 深度排序模型

排序问题是信息检索和推荐系统等领域的核心问题之一[8]，例如，搜索引擎需要将网页搜索结果按照与用户的检索目的的符合程度进行排序；推荐系统需要把候选物品按照用户可能感兴趣的程度进行排序，排序结果的精准度和合理性会直接影响检索和推荐的质量。

- 排序学习

传统的排序模型构建过程一般通过人工依据经验，去调整排序模型中所涉及到的一些参数，但这些经验参数不易调节且易产生过拟合；另一方面，尽管这些不同的排序模型大体上都使得排序效果得到了一定的性能提升，但如何将不同排序模型融合在一起以构建一个性能更优的统一排序模型，并不易于处理。同时，随着影响排序性能的排序特征的不断增加，排序特征已有成百上千种，传统的排序模型的构建方法已不再适于处理如此多维和复杂的排序特征。而机器学习方法具有能自动调整参数，融合多个模型的结果，通过正则化的方式避免过拟合等优点。在如此背景下，涌现了大量的研究者运用不同的机器学习技术去训练排序模型以解决信息检索中的排序问题，并由此产生了信息检索与机器学习交叉的一个热点研究领域--排序学习。排序学习（Learning to rank）就是利用机器学习方法在排序学习数据集上进行训练，自动产生排序模型，从而解决排序问题。和传统排序模型相比，排序学习的优势在于对众多排序特征进行组合优化，对相应的大量参数自动进行学习，最终得到一个高效精准、更加优化的排序模型[9]。

排序学习方法可以根据其训练方式分为 3 类，包括逐点训练（pointwise），成对训练（pairwise）和列表训练（listwise）。其中，逐点训练的训练目标是优化对于一个文档的相关性分数估计，大部分的回归和分类机器学习方法都能用来训练逐点训练排序学习。成对训练排序学习每一次关注两个文档，给定两个文档，该排序学习会训练给出两个文档的相对顺序，一些比较流行的成对训练排序学习方法包括 RankNet, LambdaRank and LambdaMART。列表训练排序直接对整个列表进行训练，目标为直接优化列表的相关性排序，其训练目标可以是直接优化相关性排序指标，例如 NDCG 等，也可以是最小化刻画想要关注的列表的某一特性的损失函数，例如 ListNet 和 ListMLE 等模型。

排序学习方法将机器学习方法引入到信息检索的文档相关性排序问题中，充分考虑各种排序方法对最终排序结果的影响，通过训练学习排序模型，将各种排序方法视为特征，对文档的相关性做综合的评估。排序学习是一个信息检索与机器学习相结合的研究领域。

- 深度学习

针对排序问题，传统解决方案大多依赖于人工经验，由专家根据历史数据和待排序项的特征，通过组合一系列排序规则得到排序公式。随着对排序问题研究的不断深入，目前比较常用的做法是利用机器学习相关技术解决排序问题。与传统解决方案相比，基于机器学习的排序模型具有更高的计算效率和排序准确度，得到的排序结果也具有更强的客观性。近年来，深度学习（Deep Learning）成为学术研究的热点方向，取得了一系列研究成果。深度学习算法模型与逻辑回归模型、支持向量机以及决策树类算法等传统机器学习算法模型相比，主要区别体现在深度学习模型的网络结构包含更多更深的层级，并且明确强调特征表示学习的重要性。该模型基于神经网络模型，却比简单的神经模型更为复杂，所处理的问题也更为复杂多样。最简单的深度学习模型莫过于多层感知机模型，其实深度指的就是隐层的数量，具有一个隐层的神经网络成为浅层神经网络，具有两层和两层以上的神经网络模型就可以称为深层神经网络模型也称为深度学习模型，如图 5，将传统的一次非线性变换转换为多次的非线性运算组合构成了深度学习，深度神经网络模型比传统的神经网络模型具有更强的表示能力[10]。

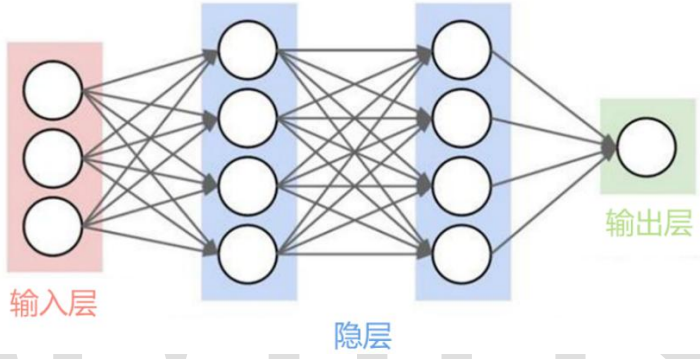


图 5 深度学习模型

- 基于深度学习的排序模型

深度排序模型（Deep Ranking Model）中比较有代表性的是神经信息检索（Neural Information Retrieval）。用于神经信息检索的排序模型使用的是浅层或深层神经网络来对搜索结果进行排序。对模型进行排名的传统学习采用有监督的机器学习技术和神经网络框架，通过人工定义的信息检索特征进行学习排序。最近提出的神经模型，可以在同一个向量空间对查询词和文档词汇之间的距离进行定量计算，距离越近，查询词与对应文档越相关。

2.1.2.2 向量空间模型

向量空间模型认识到布尔匹配太有限，提出了一套可以进行部分匹配的框架[11]。这是通过对查询和文档中的索引项赋予非二值权值实现的。这些权值最终用来计算系统中存储的文档和用户查询之间的相似度。通过对检出文档按照相似度的降序排序，向量模型考虑和查

询仅有部分匹配的文档。相比由布尔模型检出的文档集，其主要效果在于，排序的文档提供了更精准的答案更符合用户的信息需求。

向量空间模型概念简单，把对文本内容的处理简化为向量空间中的向量运算，并且它以空间上的相似度表达语义的相似度，直观易懂。当文档被表示为文档空间的向量，就可以通过计算向量之间的相似性来度量文档间的相似性。文本处理中最常用的相似性度量方式是将文档向量与查询向量量化成余弦距离。

向量空间模型主要优点是：1）权值索引提高了检索质量；2）它的部分匹配策略检出了近似于查询条件的文档；3）词组的权重不是二元的；4）允许计算文档和索引之间的连续相似程度。

但是，向量空间模型也有如下局限性：1）不适用于较长的文件，因为它的相似值不理想（过小的内积和过高的维数）；2）检索词组必须与文件中出现的词组精确匹配，不完整词组（子字串会导致“假阳性”匹配）；3）语义敏感度不佳；具有相同的语境但使用不同的词组的文件不能被关联起来，导致“假阴性匹配”；4）词组在文档中出现的顺序在向量中间中无法表示；5）假定词组在统计上是独立的，并且权重是直观上获得的而不够正式。

2.1.2.3 潜在语义索引模型

通过索引项集合来总结文档与查询内容会导致糟糕的检索质量，这咎于两点，首先，许多不相干的文档可能包含在答案集中；其次，无法检索出未被查询中的关键词索引的相关文档。造成这两点的主要原因是基于关键词集合的检索过程固有的模糊性。

概念直接是存在联系的。但是基于索引项集合来检索文档，却是基于索引项匹配而不是基于概念匹配。但是一篇文档可能与另一篇文档共享了概念。解决该问题的方法是，潜在语义索引模型。

潜在语义索引模型（Latent Semantic Indexing，LSI）的缩写，中文意译是潜在语义索引，指的是通过海量文献找出词汇之间的关系。当两个词或一组词大量出现在一个文档中时，这些词之间就可以被认为是语义相关的[12]。

潜在语义分析（Latent Semantic Analysis）或者潜在语义索引（Latent Semantic Index），是 1988 年 S.T. Dumais 等人提出了一种新的信息检索代数模型，是用于知识获取和展示的计算理论和方法，它使用统计计算的方法对大量的文本集进行分析，从而提取出词与词之间潜在的语义结构，并用这种潜在的语义结构，来表示词和文本，达到消除词之间的相关性和简化文本向量实现降维的目的。

潜在语义索引具有框架定义完整、优化准则清楚的特点，但是它也存在一些局限性，主要表现在：1）潜在语义的应用取决于具体的文档集合，比较适用于词汇异构度很高的文档集合，即文档集合中不同的文档采用不同的词汇来描述同一个概念，但是如果文档中的词汇

异构度较低，则应用潜在语义索引的效果将不太明显；2）潜在语义索引的速度比传统的向量空间方法慢，因为它需要进行高阶矩阵的运算，计算查询字段和每篇文档的相似度；3）奇异值分解存在局限性，它假设数据的分布是正态分布，然而类似词频的统计数据并不符合正态分布的条件。

2.1.3 概率模型

2.1.3.1 经典概率模型

概率模型由 Robertson 和 Sparck Jones 在 1976 年提出，他们利用了相关反馈信息逐步求精以期望获得理想的查询结果。概率模型的基本思想是：根据查询 Q 将文档集合中的文档分为两类，与 Q 相关的集合 R，与 Q 不相关的集合 R’。对于相同类的文档集，各个索引项的分布相同或相近；对于不同类的文档集，各个索引项分布不同[13]。由此可见，对文档中各个索引项的分布进行计算，依据计算出来的分布情况，我们就可以对文档和查询的相关度进行判定。

到目前为止比较常用的概率模型公式是 Robertson 提出的 BM25 公式。BM25 模型是在标准概率公式变体上经过一系列实验诞生的。这些实验是出于这样的观察，反比文档频率；项频；文档长度归一化。

BM25 算法通过加入文档权值和查询权值，拓展了二元独立模型的打分函数。这种拓展是基于概率论和实验验证的，并不是一个正式的模型。BM25 模型在二元独立模型的基础上，考虑了单词在查询中的权值以及单词在文档中的权值，拟合综合上述考虑的公式，并通过实验引入经验参数。

概率模型的主要缺点是对文本集的依赖性过强，而且条件概率值很难估计。概率模型的一个特例是贝叶斯网络，该网络以概率的方式定义了关键词的权重随着与其相关的关键词的权重的改变而改变方式。由于该模型适用于超文本信息系统，因而该模型的应用越来越广泛。但该模型仍然有它自己的缺点：其计算复杂度很大，因而并不适合很大的网络。

2.1.3.2 语言模型

语言模型在应用于信息检索之前，已经在语音识别、机器翻译及中文分词中得到了成功应用，具有准确性高、容易训练、容易维护等优点。

语言模型建模方法大致分为两类：一种是完全依赖大规模文本数据，进行统计建模；另一种是基于乔姆斯基的形式语言为基础的确定性语言模型，该建模方法更加注重语法的分析[14]。

从基本思路来说，其他检索模型都是从查询到文档进行考虑，即给定用户查询如何找出相关文档。然而，语言模型正相反，是一种逆向思维方式，即由文档到查询进行考虑，为每个文档建立不同的语言模型，判断由文档生成查询的概率是多少，根据这个概率大小进行排序作为最终搜索结果。

应用于 IR 后，语言模型和文档紧密联系，当输入查询 q 后，文档依据查询似然概率或者文档在该语言模型下能产生该查询的概率进行排序。

但语言模型面临数据稀疏问题，即查询词不在文档中出现，整个生成概率将为 0，所以语言模型引入了数据平滑，避免零概率出现。常见的平滑方式有两种：Jelinek-Mercer 平滑方法与 Dirichlet 先验的贝叶斯平滑方法。

2.1.4 其他模型

2.1.4.1 多媒体模型

随着信息技术的发展，信息的呈现方式呈现多元化趋势，信息检索也不再局限于单纯的文字检索，图像、视频等多媒体数据已经成为人们获取与传播信息的主要媒介，从各种形式的媒体源中提取判别性描述的技术问题提上了日程。

面对海量多媒体数据，如何实现快速准确的信息检索，一直是多媒体研究领域的特点问题。最早的多媒体检索研究可以追溯到 20 世纪 70 年代末期，当时主要是依赖人工标注生成媒体数据的文本标签，利用文本匹配完成检索。本世纪初，随着计算机视觉、模式识别、机器学习等技术的进步，逐步发展出多媒体内容自动标注方法，用于大规模数据的管理与检索。多媒体信息检索（Multimedia Information Retrieval，MIR）是计算机科学的研究学科，是指从多媒体数据源中提取语义信息。数据来源可以是直接可感知的媒体，比如音频、图像和视频，也可以是间接可感知的来源，比如文本、语义描述、生物信号以及不可感知的来源。多媒体检索的基本流程，如下图所示[15]：

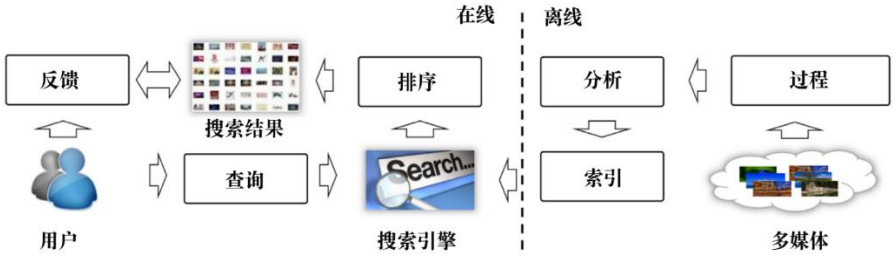


图 6 多媒体检索的基本流程

多媒体检索的研究可以分为三大类，即媒体内容特征提取技术、媒体内容表示技术和媒体内容分类技术等。

- **特征提取**

多媒体对象的庞大规模以及它们的冗余和可能的噪声是研究多媒体特征提取技术的动机。通过特征提取可以实现两个可能的目标，即媒体内容概述和通过自相关或互相关对模式进行检测。

- **内容表示**

多媒体信息检索意味着采用多个信道来理解媒体内容。每个通道都由特定于媒体的特征转换来描述。

- **内容分类**

大部分机器学习算法都可以用于多媒体内容的分类，即判断某一多媒体内容所属类别/标签。不同的方法可能适用于不同的任务，例如，Hidden Markov Model 在语音识别中是最先进的，而 Dynamic Time Warping 是基因序列比对中的最新技术。

多媒体检索经历了十几年的发展，然而检索性能的提升依然受到“意图鸿沟”与“语义鸿沟”的制约。学术界针对此问题，提出了一系列查询技术帮助用户清楚地表达检索意图以及反馈技术帮助系统准确地理解用户意图与媒体数据，有效提升了检索性能[16]。

2.1.4.2 跨语言检索模型

- 跨语言信息检索

随着互联网资源的多语言性和用户所使用语言的日益多样性，跨语言信息检索成为越来越重要的研究领域。跨语言信息检索（Cross-language Information Retrieval，CLIR）是指用户以一种语言提问，检出另一种或几种语言描述的信息资源的信息检索技术和方法。跨语言信息检索中，用户用以表达自己的信息需求，构造检索提问式的语言称为源语言（Source Language），被检索的信息资源所使用的语言称为目标语言（Target Language）。而要实现语言之间的转换，首先要使计算机能理解自然语言文本的意义，然后能以自然语言文本来表达给定的意图、思想等。例如自动识别一份文档中所有被提及的人与地点；识别文档的核心议题；在众多合同中，将各种条款与条件提取出来并制作成表；或者通过精心选定的某些特征和文本中的某些元素结合来识别一段文字，通过识别这些元素可以把某类文字同其他文字区别开来，比如垃圾邮件同正常邮件等等。跨语言信息检索是在对自然语言理解的基础之上，其关键问题是要使查询语言与文档语言在检索之前达成一致。使用户以一种语言提问，可以检索出另一种语言或多种语言描述的相关信息。例如，输入中文检索式，跨语言检索系统会返回英文、日文等语言描述的信息，而且这些信息不仅仅是文本信息，还可以是其他形式的信息[17]。

- 跨语言检索的关键技术

在跨语言检索中主要涉及的关键技术有计算机信息检索技术、机器翻译技术和歧义消解技术。计算机信息检索技术完成提问与文档之间的匹配，机器翻译技术完成不同语言之间的语义对等，歧义消解技术则解决翻译过程中的多义和歧义问题。

计算机信息检索技术。计算机信息检索技术主要是自动搜索技术、自动标引技术、语言处理技术和自动匹配技术。检索系统利用网络爬虫进行网络信息的收集，然后利用自动标引技术对搜集的信息进行标引，使用相应的语言处理技术，实现 2 种语言的相对应，形成索引数据库。用户输入检索式，计算机把检索式与数据库中的索引项进行匹配，按检索式与标引项相关度的大小排序输出检索结果。

机器翻译技术。在跨语言检索中，所要解决的问题实际上是一个语言处理问题。不同于单一语种的语言信息检索和机器翻译，也不是两种技术的简单叠加，而是一种有机的融合，有着自身的特点和专门的研究内容。机器翻译技术实质上是一种能够将一种语言的文本自动翻译成另一种语言文本的计算机程序，核心是保持两种文本（源语言文本和目标语言文本）的语义对等。由于在翻译过程中，源语言文本中的词往往对应目标语言描述的几个词，所以要选择最合适的词或相关处理以达到意义上的一致。在跨语言检索中，翻译的准确性直接决定了检索的准确性，准确性的提高需要利用自然语言处理与机器翻译相结合的技术，而由于这涉及复杂的计算机语义分析技术，因此机器翻译的效果还远未达到人们所期望的水平。

歧义消解技术。跨语言信息检索涉及到两种语言之间的相互转换，在此过程中主要会出现的问题是歧义问题[18]，需要解决自然语言文本和对话的各个层次上广泛存在的各种各样的歧义性或多义性问题。在自然语言中，一词多义和一义多词的现象是非常普遍的，对查询进行处理时，确定检索词的确切含义是非常重要的，即要把带有潜在歧义的自然语言输入转换成某种无歧义的计算机内部表示，这需要大量的知识和推理。而对被检索文献而言，要提高查准率，就需要明确文献中出现的检索词的含义，以判断其相关性。

跨语言检索的出现是为了满足网络资源语种多样性，克服用户掌握语言差异所带来的检索语言障碍。随着信息全球化的进程不断加快，人们对于跨语言信息检索的需求也越来越迫切。

2.2 信息推荐部分前沿技术

2.2.1 深度推荐模型

深度学习是机器学习领域一个重要研究方向，近年来在图像处理、自然语言理解、语音识别和在先广告等领域取得了突破性进展。将深度学习融入推荐系统中，研究如何整合海量的多源异构数据，构建更加贴合用户偏好需求的用户模型，以提高推荐系统的性能和用户满意度，成为基于深度学习的推荐系统的主要任务。

深度学习的最大优势就是能够通过一种通用的端到端的过程学习到数据的特征，自动获取到数据的高层次表示，而不依赖于人工设计特征。因此，深度学习在基于内容的推荐中主要被用于从项目的内容信息中提取项目的隐表示，以及从用户的画像信息以及历史行为数据中获取用户的隐表示，然后基于隐表示计算用户和项目的匹配度来产生推荐。在假设用户和项目携带辅助信息的情况下，深度神经网络模型被作为有效的特征提取工具。

深度学习由于能够适应于大规模数据处理，目前被广泛应用于协同过滤推荐问题中。基于深度学习的协同过滤方法主要是将用户的评分向量或项目的被评分向量作为输入，利用深度学习模型学习用户或项目的隐表示，然后利用逐点损失（point-wise loss）和成对损失（pair-wise loss）等类型的损失函数构建目标优化函数对深度学习模型的参数进行优化，最后利用学习到的隐表示进行项目推荐。

混合推荐的主要思路是融合基于内容的推荐方法与协同过滤，将用户或项目的特征学习与项目推荐过程集成到一个统一的框架中，首先利用各类深度学习模型学习用户或项目的隐特征，并结合传统的协同过滤方法构建统一的优化函数进行参数训练，然后利用训练出来的模型获取用户和项目最终的隐向量，进而实现用户的项目推荐。

基于深度学习的推荐系统通常将各类用户和项目相关的数据作为输入，利用深度学习模型学习到用户和项目的隐表示，并基于这种隐表示为用户产生项目推荐。基本的架构如图 7 所示，包含输入层、模型层和输出层[19]。

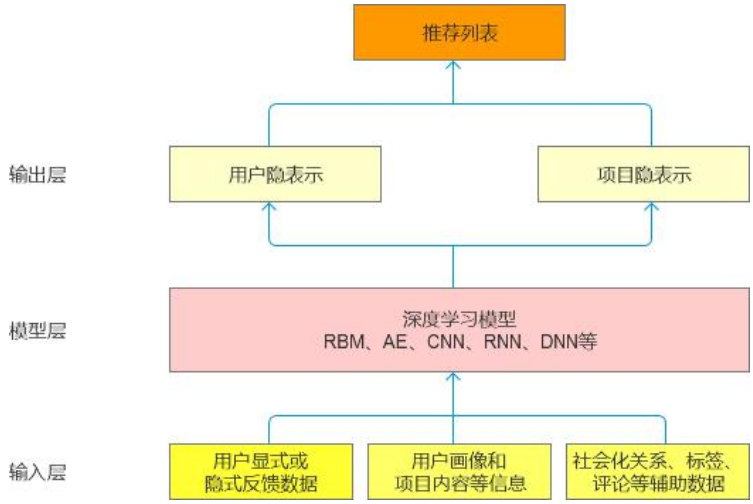


图 7 基于深度学习的推荐系统框架

输入层的数据主要包括：用户的显式反馈（评分、喜欢/不喜欢）或隐式反馈数据（浏览、点击等行为数据）、用户画像（性别、年龄、喜好等）和项目内容（文本、图像等描述或内容）数据、用户生成内容（社会化关系、标注、评论等辅助数据）。模型层使用的深度学习模型比较广泛，包括自编码器、受限玻尔兹曼机、卷积神经网络、循环神经网络等。在输出层，通过利用学习到的用户和项目隐表示，通过内积、Softmax、相似度计算等方法产生项目的推荐列表。

当前深度学习在推荐系统研究中的应用可以分为五个方向：

- 深度学习在基于内容的推荐系统中的应用。利用用户的显式反馈或隐式反馈数据、用户画像和项目内容数据，以及各种类型的用户生成内容，采用深度学习方法来学习用户与项目相似的项目推荐给用户。
- 深度学习在协同过滤中的应用。利用用户的显式反馈或隐式反馈数据，采用深度学习方法学习用户或项目的隐向量，从而基于隐向量预测用户对项目的评分或偏好。
- 深度学习在混合推荐系统中的应用。利用用户的显式反馈或隐式反馈数据、用户画像和项目内容数据，以及各种类型的用户生成内容产生推荐，模型层面主要是基于内容的推荐方法与协同过滤方法的组合。
- 深度学习在基于社会网络的推荐系统中的应用。利用用户的显式反馈或隐式反馈数据、用户的社会化关系等各类数据，采用深度学习模型重点建模用户之间的社会关系影响，更好地发现用户对项目的偏好。
- 深度学习在情景感知的推荐系统中的应用。利用用户的显式反馈或隐式反馈数据，以及用户的情境信息等各类数据，采用深度学习模型对用户情境进行建模，发现用户在特定情境下的偏好。

总的来说，基于深度学习的推荐系统研究利用深度学习方法学习用户和项目的隐表示来进行项目推荐，但是不同类型的方法在深度学习模型、数据类型以及推荐对象等方面存在着差异。

2.2.2 基于关联规则的推荐

基于关联规则的推荐（Association Rule-based Recommendation）是以关联规则理论为基础，把已购商品作为规则头，规则体为推荐对象[20]。

关联规则是用于在海量数据中挖掘出其背后隐藏的事务项之间的联系，通过数据之间的联系中的有用内容来获取更大的利益。它是用来表示两个项目事务之间的依赖性，假如两个项目事务之间存在一定的关联，那么其中一个项目事务可以以一定的概率作为前提条件来推断另一个项目事务。其挖掘目标是为了从庞大的数据集中发现项之间不易得到的关联。它的应用由最开始的购物篮分析发展到分类关联分析、知识提炼和推荐、蛋白质内部成分分析、软件故障挖掘，机器故障推断、交通事故模式分析等。关联规则的理论研究趋势是由最初的开频繁模式发展到闭合形式挖掘、增量形式挖掘、比较兴趣度、流体数据等不同表达方式的关联规则[21]。联规则算法得到充分的应用，尤其是在电子商务领域，通过关联规则解决了用户对大量信息的处理问题，给用户提供个性化的推荐，不仅节约了用户的宝贵时间，同时提高了整个网站的营业额。

基于关联规则的推荐技术重点在于其关注用户行为之间的关联模式，比如一个用户买了一袋面包，在大多数情况下该用户还会选择再买一袋牛奶共同作为早餐，因此可以在面包和牛奶之间建立关联关系，从而根据这种关联关系推荐其他的产品。另外，比较著名的关联效应是“啤酒和尿布”的案例，年轻父亲在超市购物时会根据需求购买尿布，再考虑到年轻父亲喜欢喝啤酒的习性，如果在尿布旁边摆放啤酒的话，他有很大几率会选择同时购买啤酒，这样看起来毫无关联的啤酒和尿布之间就有了一定的关联关系，这种关联性不只体现在实体超市，在电子商务网站中也很突出。在电子商务网站购物时，也许年轻父亲不是专门去买啤酒的，但是如果其在将必要的婴儿尿布放入购物车后，系统能根据该用户的年龄特征和喜好，来推荐类似于啤酒之类产品的话，能促进用户的交叉购买，这种基于关联性的交叉销售行为不仅能给用户带来了方便，还可以提高网站的销售量。

相对于其他推荐系统，基于关联规则的推荐系统具有自己的优势：1）数据源相对简单，不需要特殊的数据源，只需要有准确的交易记录即可；2）对用户购买行为具有的预测的能力，而且能够挖掘用户的潜在兴趣；3）可以对不同类型的商品进行挖掘，也就说对商品特性没有特殊的要求[22]。

2.2.3 协同过滤推荐

协同过滤技术是推荐系统中最为成功的技术之一，被广泛用于预测用户兴趣偏好的应用领域。在日常生活中，我们在选择商品时往往会向好朋友咨询意见，从而帮助我们做出决策。协同过滤正是把这一思想运用到个性化推荐中来，即基于兴趣爱好相似的用户对某些项目的评价来向目标用户推荐合适的项目[23]。协同过滤机制的主要目的在于根据已有数据之间的关系，计算用户之间的相似度，找到有共同兴趣爱好的用户，从而产生推荐。例如，如果有两个用户对某些商品的评分相似，则系统认为这两个用户的偏好是相似的，因此会将一个用户评价较好的商品推荐给另一个用户。

基于协同过滤的推荐技术并不关心用户信息或者商品项目信息，而是通过对目标用户的历史行为，主要是对商品的历史评分数据进行分析，找到和目标用户兴趣爱好相似的用户群体，依据这些用户对商品做出的评价来预测目标用户对未评分商品的评分，然后向目标用户推荐合适的商品集。

协同过滤推荐技术有如下优点：1）适用于复杂的非结构化的数据，如电影、音乐等数据，多媒体等资源的内容特征分析难度较大，但是协同过滤技术利用的数据易于提取和表示，例如：用户评分、购买记录、浏览记录等；2）善于发现用户新的兴趣点，推荐过程中相似用户的“建议”能够拓宽推荐关注点，可以推荐和用户以往喜欢的项目完全不同的事物，即发现用户可能喜欢但未曾察觉的事物，不需要专业领域的知识；3）智能性，协同过滤技术不需要用户自己定位兴趣点，例如填写调查问卷等，而是自动根据用户的历史评分信息等显式信息或浏览信息等隐式信息为用户做出相应的推荐。

2.2.4 基于内容的推荐

基于内容的推荐技术起源于信息检索和信息过滤，它根据商品项目的内容信息和用户的偏好之间的相关性来向用户推荐信息。基于内容的推荐技术通常基于这样的假设，拥有相似特征的商品会得到目标用户相似的评分。例如，用户喜欢一部关于战争和爱情的电影，他很有可能对其他与战争和爱情有关的电影也感兴趣。基于内容的推荐技术一般通过类别或特征标签选择来获取用户的需求和喜好[23]。基于内容的推荐技术主要是通过信息过滤来获取更有价值的信息，这些信息主要包括项目的特征信息和项目的描述信息，它不需要获取用户对项目的行为数据，而是通过各种方法对项目的特征属性进行定义。

基于内容的推荐技术首先分析用户已经评价项目的属性来定位用户的兴趣偏好，再通过比较用户兴趣点与项目之间的相似性来为用户产生推荐。用户的兴趣模型常取决于所用的学习方法，比较常见的有决策树、神经网络、贝叶斯分类器、聚类等。基于内容的推荐技术的

最关键之处在于对项目的理解程度，它需要从项目中抽取可以代表项目的典型特征词，对项目结构进行分析，然后构建项目的信息模型。目前大部分基于内容的推荐技术通常是对文本信息进行研究。

在对电影进行个性化推荐时，该算法首先根据用户的历史评分记录，对用户评分较高电影的某些共同属性（比如演员、电影类别等）进行分析，然后搜索与这些属性总体相似的其他电影，并推荐给用户。基于内容的推荐技术是以产品为核心，由于产品的属性基本上不会发生很大的变化，故而产品间的关系也相对稳定，那么基于内容的推荐技术也是比较稳定的，具有普遍适用性。

基于内容的推荐技术有如下优点：1）推荐结果直观，可解释性好，推荐给用户的项目的内容特征和用户评分较高的项目的内容特征具有很强的相似性，用户容易接受，从而使用户对基于内容的个性化推荐的认可度较高；2）在一定程度上能解决新项目的问题（即项目冷启动问题），当一些项目新加入到推荐系统中时，该算法能够利用这些新项目的内容特征和用户偏好做匹配，其被推荐的可能和老项目是相同的。而且该算法不会受到评分稀疏性问题的影响，能够将新产品和非流行产品及时推荐给用户。

2.2.5 组合推荐方法

每种推荐方法都有各自的优缺点，例如，协同过滤算法在面对复杂的非结构化项目时占据优势，但很难为拥有特殊兴趣的用户推荐优质的项目;而基于内容的推荐可以为拥有特殊兴趣的用户给出推荐，但又不能发现用户的潜在兴趣。所以有很多学者提出将多个推荐方法混合在一起，从而达到更好的推荐效果，也就是使用组合推荐（Hybrid Recommendation），目前研究和应用最多的是基于内容推荐和协同过滤推荐的组合不同的组合思路适用于不同的应用场景[24]，这里简要介绍几种比较流行的组合方法：

- 加权组合（Weighted Hybridization）：就是将多种推荐算法的推荐结果进行加权混合。最简单的方式是线性组合，将几种不同的推荐算法的结果按照一定权重组合起来，但该方法的权重的设置是一个难题，需要根据实际情况，反复实验而定，从而达到最好的组合效果。这种组合方式的前提是，对于整个空间中所有可能的项，使用不同技术的相关参数值都基本相同。
- 切换组合（Switching）：不同的推荐技术适应于不同的需求，切换组合就是根据不同的推荐场景切换使用不同的推荐算法。例如，先使用基于内容的推荐算法，如果它不能产生质量较高的推荐，再尝试使用协同过滤算法。
- 混合式组合（Mixed）：就是使用多种推荐技术给出多种推荐结果，供用户参考。例如，可以根据用户偏好记录构建的兴趣模型，采用基于内容的推荐，向用户推荐相似项目；

同时，使用协同过滤算法为活动用户预测潜在可能感兴趣项目。最后将两种推荐结果都推荐给用户。其实，Amazon、当当网等很多电子商务网站都是采用这样的方式，用户可以得到很全面的推荐，也更容易找到他们想要的东西。

- 特征组合（Feature combination）：这是最常用的组合方法之一，它组合来自不同推荐方法的数据信息，一种推荐算法产生的数据信息被另一种推荐算法所采用。例如将协同过滤的产生信息作为增加的特征向量，然后在数据集上采用基于内容的推荐算法。
- 瀑布式组合（Cascade）：这是一个分段的推荐过程，先用一种推荐方法产生一种粗糖的候选推荐结果，然后，使用另一种推荐方法在此候选推荐结果上进一步做出更精确的推荐。这样综合使用各个推荐算法的优点，得到更加准确的推荐。
- 特征扩充（Feature）：将一种推荐算法产生的特征信息嵌入到另一种推荐算法中。例如，使用聚类方法作为关联规则方法的预处理步骤，先对评分矩阵进行聚类，再针对每个类簇进行关联规则挖掘。根据用户的当时访问路径与各类簇的匹配度，确认所属类簇，再使用该类簇对应的关联规则进行推荐。
- 元层次组合（Metal-level）：用一种推荐算法产生的模型作为另一种推荐方法的输入。例如，组合基于用户的协同过滤算法和基于项目的协同过滤算法，先计算出待预测项目的相似项目集，在该相似项目集上再采用基于用户的协同过滤算法。这时计算出的用户间相似度，能较好地处理用户多兴趣问题。

2.2.6 基于知识的推荐

协同过滤和基于内容的推荐在很多情况下无法发挥作用，例如：1）有些物品我们并不会频繁购买，比如房屋，纯粹的协同过滤系统会由于评分数据很少而效果不好；2）时间跨度因素的作用很重要，多年前对物品的评分对基于内容推荐来说就不太合适，因为用户偏好会随着生活方式或家庭情况的变化而改变；3）在一些复杂的产品领域，用户希望明确定义他们的需求，例如“汽车的最高价是 x，颜色是黑色”，这种需求的形式化处理并不是纯粹协同过滤和基于内容推荐所擅长的（关于这个第三条，在以前对于推荐系统的概念中，没有把这种明确需求的形式认为是推荐系统的一种，而是认为这是一种检索系统，后面的讨论中会在做说明）。

基于知识的推荐（Knowledge-based Recommendation，KB）是一种特定类型的推荐系统，它借助于领域本体，表达语义知识，增加了项目之间的关联信息；通过领域本体中结合点、边、深度和密度对相似性计算的不同影响，算法结合信息论中的互信息相关概念，对相似性计算公式进行改进，提高了运算精度[5]。基于知识的推荐系统可以解决上述问题：1）由于不需要评分数据就能推荐，也就不存在启动问题；2）推荐结果不依赖单个用户评分，要么

是以用户需求与产品之间的相似度的形式，要么是根据明确的推荐规则；3）关于推荐系统是什么，传统解释一般强调信息过滤这一方面，即过滤出某个用户可能感兴趣的商品，而基于知识的推荐交互性很强，使得推荐系统不在仅仅被看作是一种推荐系统，而是“以一种个性化方法引导用户在大量潜在候选项中找到感兴趣或有用物品，或者将这些物品作为输出结果”的系统。

基于知识的推荐系统大致工作流程是用户指定需求，然后系统设法给出解决方案。如果找不到解决方案，用户需要修改需求（再次说明基于知识推荐系统的交互性很强）。此外，系统还要给出推荐物品的解释。基于知识推荐系统主要包括两种类型，基于约束推荐和基于实例推荐，它们的区别在于如何使用所提供的知识：基于实例的推荐系统着重于根据不同的相似度衡量方法检索出相似的物品（也就是根据相似度衡量标准检索哪些与特定用户需求相似的物品），基于约束的推荐系统以来明确定义的推荐规则集合[25]（在符合推荐规则的所有物品集合中搜索得出要推荐的物品集合）。

2.2.7 可解释性推荐

可解释性推荐是解决原因问题的个性化推荐算法，它们不仅为用户提供建议，还提供解释，使用户或系统设计人员了解推荐此类项目的原因。通过这种方式，它有助于提高推荐系统的有效性、效率、说服力和用户满意度。近年来，在现实世界的系统中已经采用了大量可解释的推荐方法，特别是基于模型的可解释推荐算法。

为了突出整个推荐系统研究中可解释推荐的位置，我们将大多数现有的个性化推荐研究分类为广泛的概念分类。具体而言，许多推荐研究任务可以被分类为解决 5W 问题-何时、何地、谁、什么、为什么，以及五个 W 通常对应于时间感知推荐、基于位置的推荐、社会推荐、应用意识推荐和可解释的推荐。

可解释的推荐研究可以追溯到个性化推荐研究中的一些最早期的工作。例如 Herlocker 等人提到，推荐系统将用于用户解释产品是什么类型的产品，例如“您正在查看的此产品与您过去喜欢的其他产品类似”，这是基于项目的协同过滤的基本思想。早期方法主要集中在基于内容的推荐或基于协同过滤的推荐上。

为了使个性化推荐模型直观易懂，研究人员越来越多地转向对可推荐模型的研究，其中推荐算法不仅提供推荐列表作为输出，而且自然地在可解释的工作中起作用。

从广泛的意义上说，人工智能系统的可解释性已经成为 20 世纪 80 年代“old”或逻辑人工智能时代的核心讨论，早期基于知识的系统预测（或诊断）很好，但无法解释原因。近年来，越来越多的研究人员意识到可解释人工智能的重要性，意在解决人工智能解释中的各种问题：深度学习、计算机视觉、自动驾驶系统和自然语言处理任务等。

作为人工智能领域的一个重要分支，可解释推荐系统（explainable recommendation）也广泛运用到生活的各个领域中[26]。根据不同的实际应用场景，推荐系统的解释有不同的形式，如基于协同的解释，基于内容的解释，基于知识和自然语言的解释和基于人口统计的解释。

2.3 信息检索与推荐领域相关资源

关于信息检索与推荐领域相关的图书、文章、研究中心、竞赛等资源整理如表 1 所示，广大读者朋友可根据自身兴趣关注了解。

表 1 信息检索与推荐领域相关资源

书籍	
书名	作者
《Introduction to Information Retrieval》	C.D. Manning, P. Raghavan, H. Schütze.
《Modern Information Retrieval》	R. Baeza-Yates, B. Ribeiro-Neto.
《Information Retrieval: Algorithms and Heuristics》	D.A. Grossman, O. Frieder.
《Managing Gigabytes》	I.H. Witten, A. Moffat, T.C. Bell.
《Finding Out About》	R. Belew.
《Information Retrieval: A Health and Biomedical Perspective》	W.R. Hersh.
《TREC: Experiment and Evaluation in Information Retrieval》	E.M. Voorhees, D.K. Harman.
《Language Modeling for Information Retrieval》	W.B. Croft, J. Lafferty.
《Readings in Information Retrieval》	K. Sparck Jones, P. Willett.
《Information Storage and Retrieval Systems》	G. Kowalski, M.T. Maybury.
《The Geometry of Information Retrieval》	C.J. van Risjbergen.
《Introduction to Modern Information Retrieval》	G.G. Chowdhury.
《Text Information Retrieval Systems》	C.T. Meadow, B.R. Boyce, D.H. Kraft, C.L. Barry.

热门文章	
文章标题	作者
《Information Retrieval》	Wikipedia
《Modern Information Retrieval: A Brief Overview》	A. Singhal
《How Google Finds Your Needle in the Web's Haystack》	D. Austin
《Simple, proven approaches to text retrieval》	S.E. Robertson, K. Sparck Jones
《What Do People Want From IR》	Bruce Croft
《The Seven Ages of Information Retrieval》	Michael Lesk
研究中心	
CMU (LTI)	卡内基梅隆大学语言技术研究所
Glasgow	格拉斯哥大学
Helsinki Institute for Information Technology	赫尔辛基信息技术研究所
IBM	国际商业机器公司
Illinois Institute of Technology	伊利诺理工大学
Microsoft Research	微软研究院
Peking	北京大学
Pittsburgh	匹兹堡大学
Queen Mary	伦敦玛丽女王大学
Sheffield	谢菲尔德大学
UIUC	伊利诺伊大学厄巴纳-香槟分校
UMASS	马萨诸塞大学

竞赛	
SIGIR	
杰拉德·索尔顿奖	http://sigir.org/awards/gerard-salton-awards/
最佳论文奖	http://sigir.org/awards/best-paper-awards/
最佳学生论文奖	http://sigir.org/awards/best-student-paper-awards/
最佳短篇论文奖	http://sigir.org/awards/best-short-paper-awards/
博士协会奖	http://sigir.org/awards/doctoral-consortium-awards/
TREC	
论文集	https://trec.nist.gov/proceedings/proceedings.html
数据信息	https://trec.nist.gov/data.html
RecSys Challenge	
获奖者	https://recsys.acm.org/challenges/
最佳论文奖	https://recsys.acm.org/best-papers/

3 人才篇

在大数据时代，信息检索与推荐技术不断迭代更新，应用范围不断扩大，众多该领域专家学者们在信息检索与推荐道路上不断探索与研究。本章节通过 AMiner 大数据平台对信息检索与推荐领域的顶级学术会议及期刊论文进行挖掘，获得这些会议和期刊最近 10 年（2009-2018 年）的论文，之后通过论文作者挖掘出该领域的专家学者，提取论文中所有学者信息，从中选出 h-index 排名最靠前的 2000 位领域活跃学者，分析了学者的分布、迁徙等情况，介绍了部分该领域国内外知名度较高的活跃学者。

我们所研究的期刊和会议包括：

TOIS（ACM Transactions on Information Systems）
SIGIR（Special Interest Group on Information Retrieval）
IPM（Information Processing and Management）
IS（Information Systems）
TWEB（ACM Transactions on the Web）
IR（Information Retrieval）
RecSys（ACM Conference on Recommender Systems）
WSDM（International Conference on Web Search and Data Mining）
ECIR（European Conference on Information Retrieval）
ICTIR（ACM SIGIR International Conference on the Theory of Information Retrieval）

3.1 学者情况概览

3.1.1 全球学者概况

学者分布地图对于进行学者调查、分析各地区竞争力现况尤为重要，图 8 为信息检索与推荐领域全球顶尖学者分布状况：

图 8 信息检索与推荐全球顶尖学者分布

由上图可以看出，从国家角度来看，信息检索与推荐领域的人才在美国最多，中国次之，英国和德国等国家也有较多的人才分布；从地区角度来看，西欧及美国东部的人才较为集中，亚洲人才主要集中在中国大东部地区，南美及非洲地区的人才非常匮乏。

在性别比例方面，信息检索与推荐领域中男性学者占比 91.4%，女性学者占比 8.6%，男性学者占比远高于女性学者（如图 9 所示）。

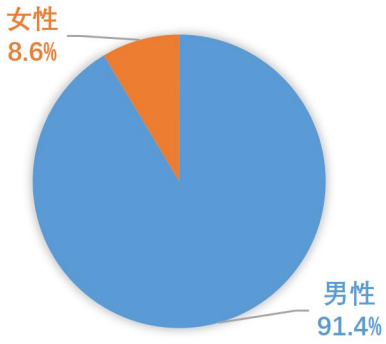


图 9 信息检索与推荐顶尖学者性别比例

信息检索与推荐领域顶尖学者的 h-index 分布如图 10 所示，分布情况整体呈阶梯状，大部分学者的 h-index 分布在中低区域，其中 h-index 在<10 的区间人数最多，有 985 人，占比 49.2%，50-60 区间人数最少，有 46 人，占比 2.3%。

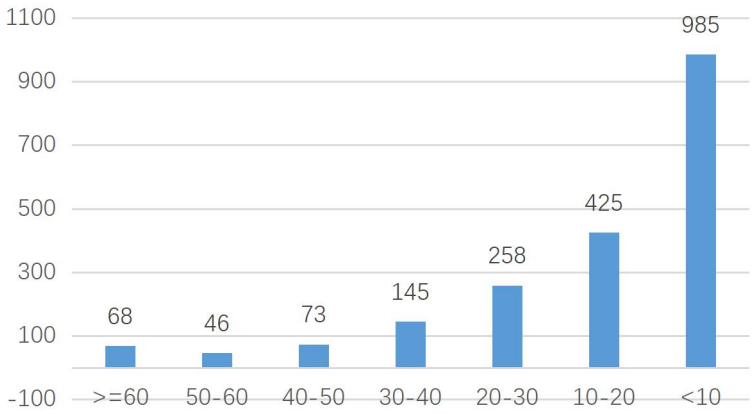


图 10 信息检索与推荐顶尖学者 h-index 分布

AMiner 可以对信息检索与推荐领域的学者的迁徙路径进行分析，如图 11 所示。从中可以看出，美国信息检索与推荐领域人才的流失和引进相对比较均衡，作为信息检索与推荐领域人才流动大国，人才输入和输出都大幅度领先，且从数据来看人才流入大于人才流出。中国、英国和德国都落后于美国，且有轻微的人才流失现象。

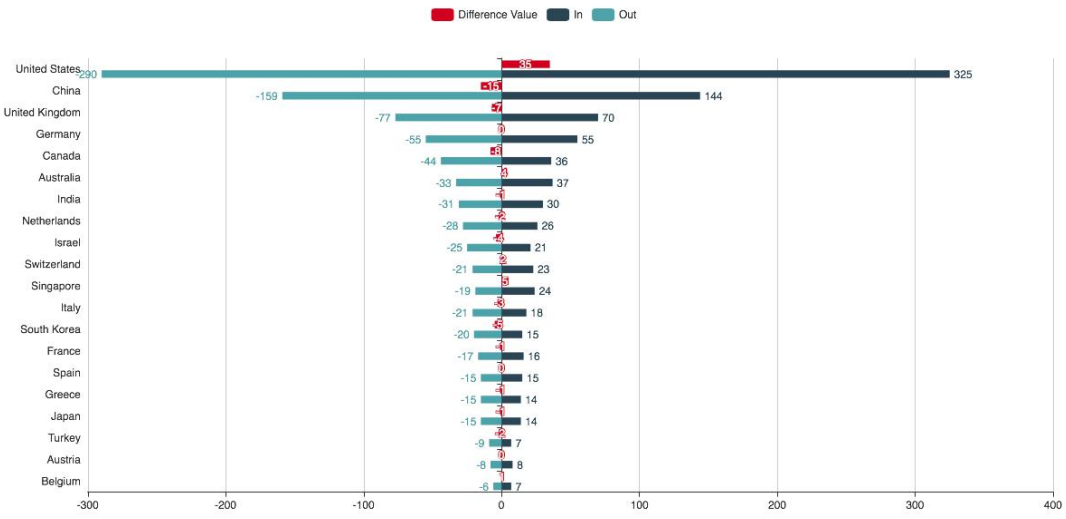


图 11 信息检索与推荐全球学者迁徙图

中国与其他国家在信息检索与推荐领域的合作情况也可以根据 AMiner 数据平台分析得到，通过统计论文中作者的单位信息，将作者映射到各个国家中，进而统计中国与各国之间合作的论文发表数量、论文引用数量，并按照合作论文发表数量从高到低进行了排序，如表 2 所示。最后选取了合作论文数量前 10 的合作关系生成了中国与各国合作论文情况对比图，如图 12 所示。

表 2 信息检索与推荐领域中国与各国合作论文情况

合作国家	论文数	引用数	平均引用数
中国-美国	1308	55827	43
中国-澳大利亚	187	3470	19
中国-英国	177	3529	20
中国-新加坡	167	6952	42
中国-加拿大	159	4420	28
中国-日本	110	1729	16
中国-法国	70	1064	15
中国-德国	64	852	13
中国-瑞士	25	348	14
中国-荷兰	22	460	21

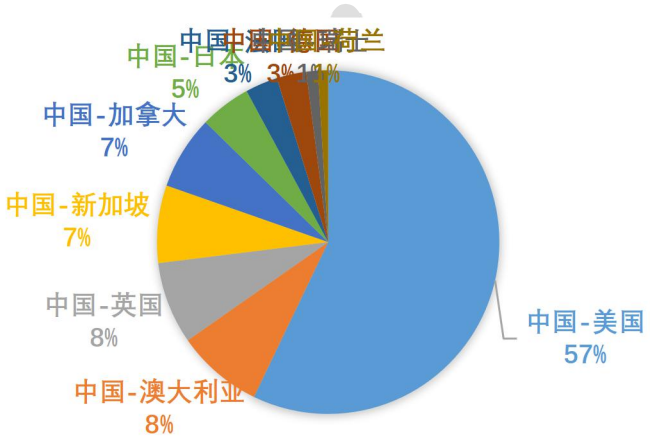


图 12 信息检索与推荐领域中国与各国合作论文情况对比图

从上面图表中数据可以看出，中美合作论文数量最多，超过前 10 名合作国家论文总和半数以上，但是与欧洲的合作较少；论文引用情况与论文数情况在趋势大体相符，其中，中国与新加坡合作的论文数虽然不是最多，但是平均引用数却与中美合作情况相近，说明在合作质量上，中-新合作达到了较高的水平。

3.1.2 国内学者分布

AMiner 选取信息检索与推荐国内专家学者绘制了学者国内分布地图，如图 13 所示。通过下图我们可以发现，京津地区在信息检索与推荐领域的人才数量最多，东部及南部沿海地区的也有较多的人才分布，相比之下，内陆地区信息检索与推荐产业人才较为匮乏，这也从

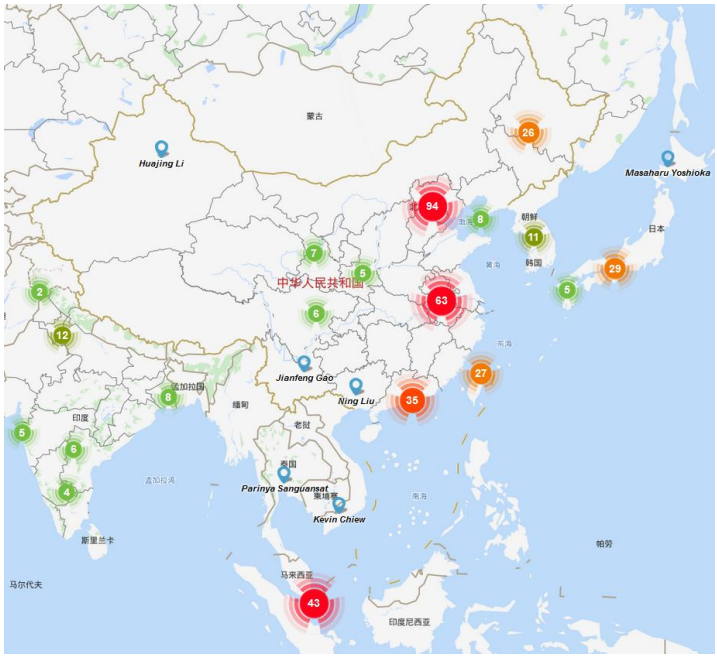


图 13 信息检索与推荐国内顶尖学者分布

一定程度上说明了信息检索与推荐领域的发展与该地区的地理位置和经济水平都是息息相关的。同时，通过观察中国周边国家的学者数量情况，特别是与日本、东南亚等亚洲国家相比，中国在信息检索与推荐领域顶尖学者数量方面具有较为明显的优势。下图是我国信息检索与推荐领域顶尖学者最多的 10 个省份：

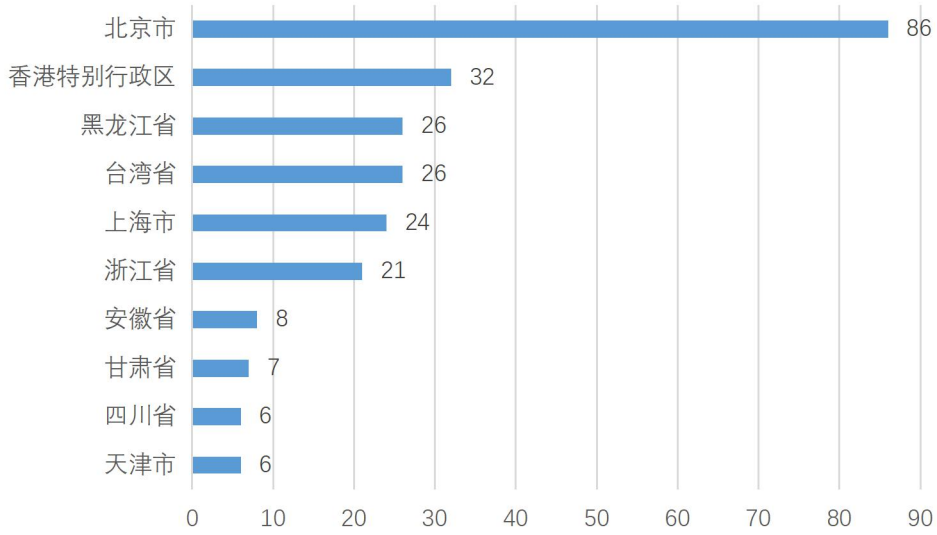


图 14 信息检索与推荐顶尖学者分布国内省份 TOP10

通过在 AMiner 信息检索与推荐人才库抓取统计的分析数据，可获得国内学者信息，可以清楚看到学者姓名、照片、职位、机构、论文数量、论文引用量及研究领域等，下图为 h-index 排名前十的国内学者：

马维英(Weiying Ma)

h-index：97 | 论文数：410 | 引用数：40208

Vice President

Today's headlines

网页 | 图像检索 | 信息检索 | 搜索引擎 | 指标 | 数据挖掘 | 移动设备 | 特征提取

5693

张磊(Lei Zhang)

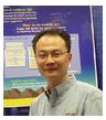
h-index：95 | 论文数：555 | 引用数：40878

Principal Research Manager

Department of Computing, The Hong Kong Polytechnic University

特征提取 | 图像检索 | 人脸识别 | 指标 | 图像标注 | 图像分类 | 搜索引擎 | 稀疏表示

1216



李航(Hang Li)

h-index：65 | 论文数：278 | 引用数：17063

Researcher

字节跳动

信息检索 | 学习排序 | 机器学习 | 网络检索 | 数据挖掘 | 搜索引擎 | 损失函数 | 文献检索

3947

俞勇(Yong Yu)

h-index：57 | 论文数：306 | 引用数：14441

Professor

上海交通大学

网页 | 语义网 | 搜索引擎 | 机器学习 | 协同过滤 | 推荐系统 | 社会化标注 | 信息检索

3646

金國慶(Irwin King)

h-index：56 | 论文数：221 | 引用数：12312

Professor

香港中文大学

数据挖掘 | 推荐系统 | 社会网络 | 机器学习 | 推荐系统 | 协同过滤 | 支持向量机

825

Zheng Chen

h-index：53 | 论文数：190 | 引用数：11216

Senior Researcher

Microsoft Research Asia

web页面 | 搜索引擎 | 网页 | 信息检索 | 机器学习 | 数据挖掘 | 链接分析

3597

刘铁岩(Tieyan Liu)

h-index：51 | 论文数：210 | 引用数：12056

Assistant Managing Director

微软亚洲研究院

学习排序 | 信息检索 | 搜索引擎 | 损失函数 | 网页 | 运动估计 | 机器学习 | 网络检索

814

文继荣(JiRong Wen)

h-index：47 | 论文数：232 | 引用数：12322

教授、院长

中国人民大学信息学院

网页 | 搜索引擎 | 信息检索 | 指标 | 信息提取 | 概率模型 | 网络检索 | 条件随机域模型

567

程学旗(Xueqi Cheng)

h-index：45 | 论文数：548 | 引用数：8757

博导

中国科学院

信息检索 | 网际网 | 数据挖掘 | 聚丙烯 | 特征选择 | 网页 | 社会网络 | 搜索引擎

3196

万小军(Xiaojun Wan)

h-index：30 | 论文数：192 | 引用数：4475

Professor

北京大学

信息检索 | 自然语言处理 | 语篇分析 | 网际网 | 单文档

423

图 15 信息检索与推荐国内学者 TOP10

3.2 典型学者

我们在选取的期刊会议中，对所涉学者及其论文关键信息进行抽取。依据年龄将抽取的学者分为资深学者、中青年学者两部分进行介绍。

3.2.1 资深学者

- Chengxiang Zhai（翟成祥）

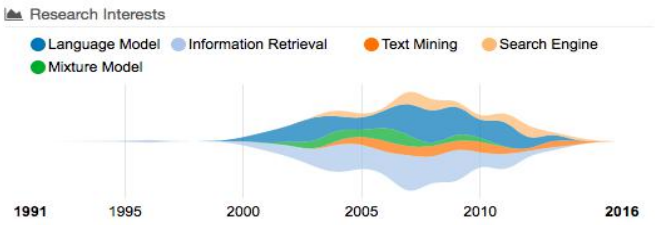
Chengxiang Zhai

H 79 A 159.89 S 91.38 C 28675 P 409

Professor

Department of Computer Science,University of Illinois at Urbana-Champaign

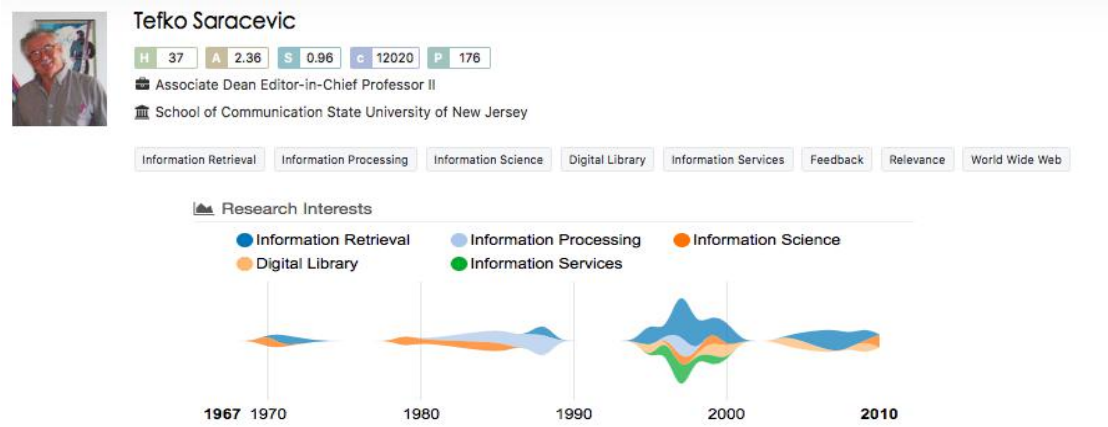
Language Model | Information Retrieval | Text Mining | Search Engine | Mixture Model | Topic Model | Probabilistic Model | Query Expansion



翟成祥是伊利诺伊大学厄巴纳-香槟分校（University of illinois at Urbana-Champaign）的教授，于 1990 年、2002 年分别获得南京大学计算机博士学位和卡耐基梅隆大学语言和信息技术博士学位，2017 年入选 ACM Fellow，曾获多项奖项：三次获得 Association for Computing Machinery SIGIR Test of Time Paper Award，2004 年度美国青年科学家和工程师最高荣誉总统奖（PECASE），Alfred P. Sloan 研究奖，IBM 优秀教师奖，HP 创新研究奖和 UIUC 优秀教师称号。

翟成祥的研究领域包括：信息检索、文本挖掘、自然语言处理、机器学习和生物信息学等。他的高引论文“*A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval*”提到信息检索的语言建模方法具有吸引力和前景，因为它们将检索问题与语言模型估计问题联系起来，语言模型估计已在语音识别等其他应用领域得到广泛研究。这些方法的基本思想是估计每个文档的语言模型，然后根据估计的语言模型查询的可能性对文档进行排名。

- Tefko Saracevic

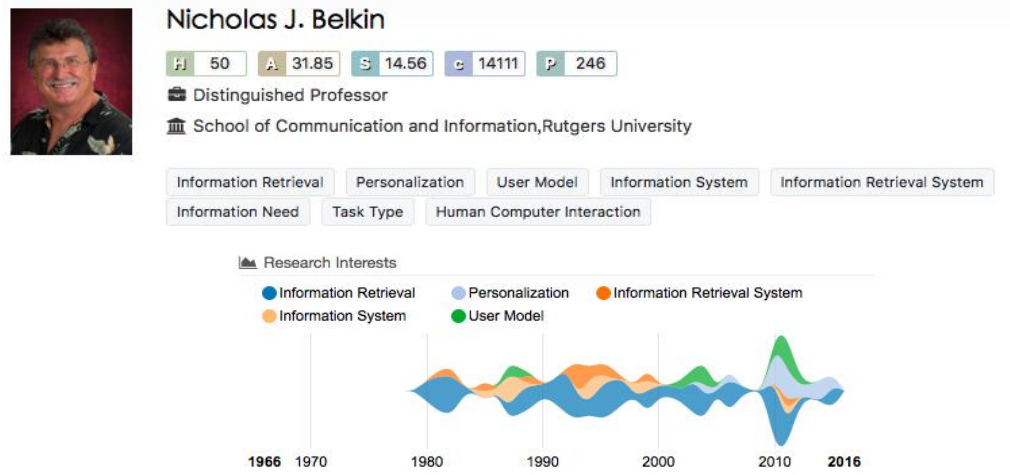


Tefko Saracevic 是罗格斯大学（Rutgers University）的教授。他在克罗地亚萨格勒布大学（University of Zagred）学习电气工程，并取得了硕士学位和博士学位。1970 年，Tefko Saracevic 在俄亥俄州克利夫兰的凯斯西储大学（Case Western Reserve University）进行信息科学研究。

Tefko Saracevic 活跃于众多专业协会，他获得了 1997 年 SIGIR / ACM 的 Gerard Salton Award；担任了 1991 年美国信息科学学会（American Society for Information Science，ASIS）的主席；获得了 1989 年美国信息科学学会期刊（Journal of the American Society for Information Science）的最佳论文奖。

Tefko Saracevic 的研究领域包括：信息检索、人机交互等。他的高引论文“*Real life, real users, and real needs: a study and analysis of user queries on the web*”，分析了由主要互联网搜索服务 Excite 的 18113 名用户提出的 51473 个查询的交易日志。将分析的重点从查询转移到用户，以深入了解 Web 用户的特征，同时进行了故障分析。

- Nicholas J. Belkin

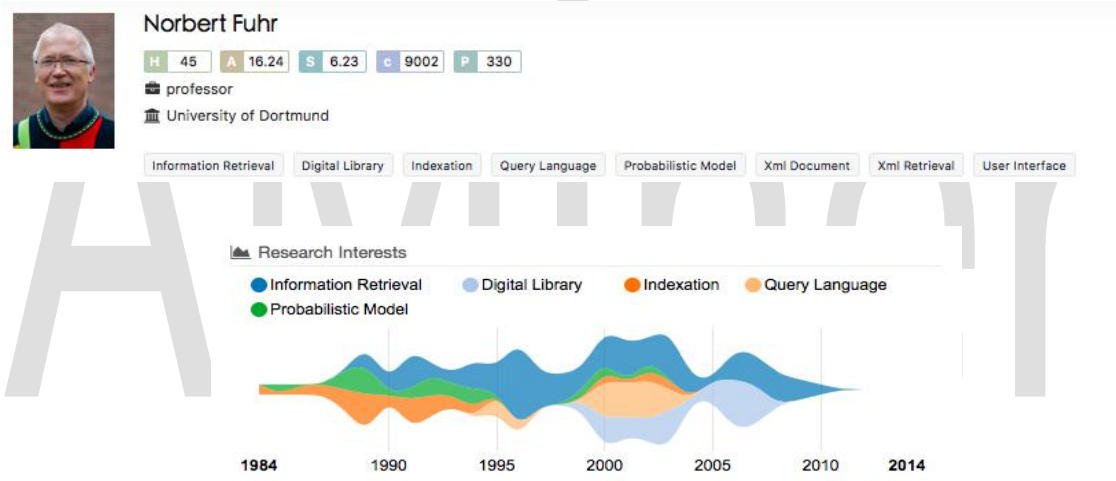


Nicholas J. Belkin 是罗格斯大学（Rutgers University）传播与信息学院的特聘教授。他取得了华盛顿大学（University of Washington）信息科学的硕士学位和伦敦大学（University of London）信息研究的博士学位。

Nicholas J. Belkin 担任了 1995-1999 年的 SIGIR 主席，2005 年美国信息科学与技术学会（Association for Information Science and Technology，ASISI&T）的会长，获得了 2015 年的 Gerard Salton 奖。

Nicholas J. Belkin 的研究领域包括：人与信息的互动、交互式信息检索、信息系统中的人机交互等。他的高引论文“*Information filtering and information retrieval: Two sides of the same coin？*”提到信息过滤是一个名称，用于描述涉及向需要的人传递信息的各种过程。虽然这个术语经常出现在描述诸如电子邮件，多媒体分布式系统和电子办公文档等应用程序的流行和技术文章中，但过滤和相关过程（如检索、路由、分类和提取）之间的区别通常并不重要。

- Norbert Fuhr

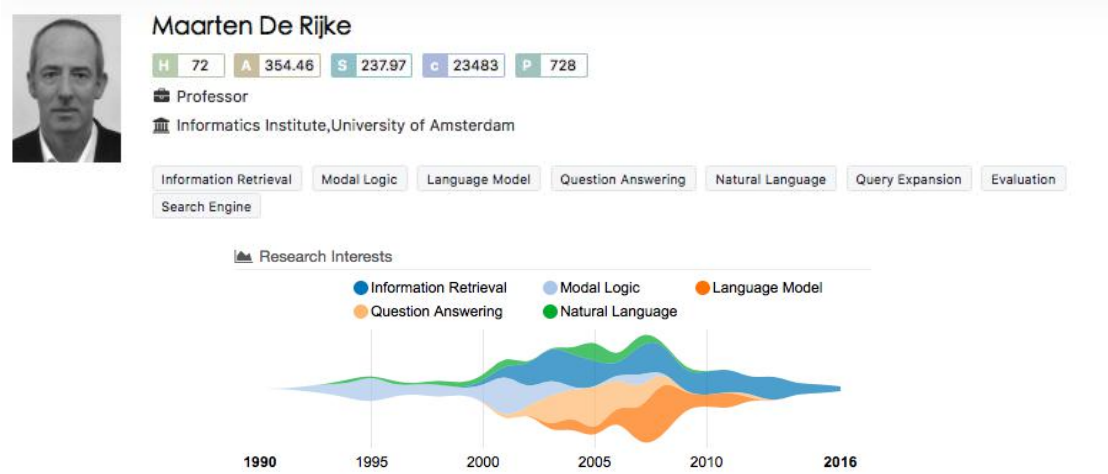


Norbert Fuhr 是德国杜伊斯堡—埃森大学（University of Duisburg-Essen，Germany）信息工程组的计算机科学教授。他取得了达姆施塔特工业大学（Technical University of Darmstadt）计算机科学的博士学位。

Norbert Fuhr 他获得过很多奖项，例如 1987 年德国文献学会（the German Society of Documentation）颁发的“Gerhard Pietsch Award”，2012 年的 Gerard Salton Award。

Norbert Fuhr 的研究领域包括：信息检索、数字图书馆等。他的高引论文“*A probabilistic relational algebra for the integration of information retrieval and database systems*”，提出了概率关系代数（PRA）。在 PRA 中，元组被赋予概率权重，同时还展示了扩展语义产生相同结果的表达式。

- Maarten de Rijke

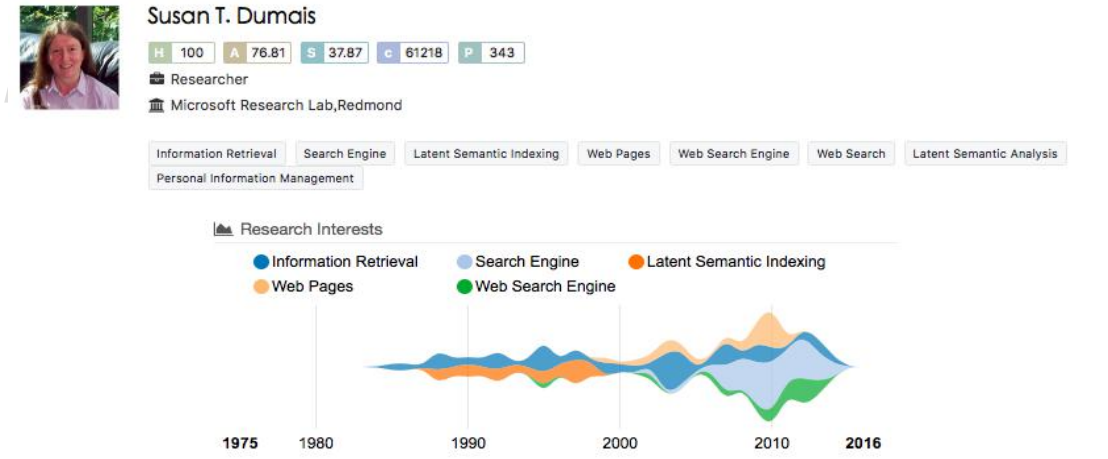


Maarten de Rijke 是荷兰的计算机科学家。他于 1998 年加入阿姆斯特丹大学，并在 2004 年被任命为阿姆斯特丹大学信息学院教授，领导阿姆斯特丹大学信息与语言处理小组。

2017 年，Maarten de Rijke 当选为荷兰皇家艺术与科学院（Royal Netherlands Academy of Arts and Sciences）院士。

Maarten de Rijke 最初的工作侧重于模态逻辑和知识表示，但自 21 世纪初以来，他主要从事信息检索工作。他的高引论文“*Expertise retrieval*”提到，作为信息检索领域新兴的分支学科，专业知识检索已经引起了极大的兴趣和丰富的成果。

- Susan Dumais

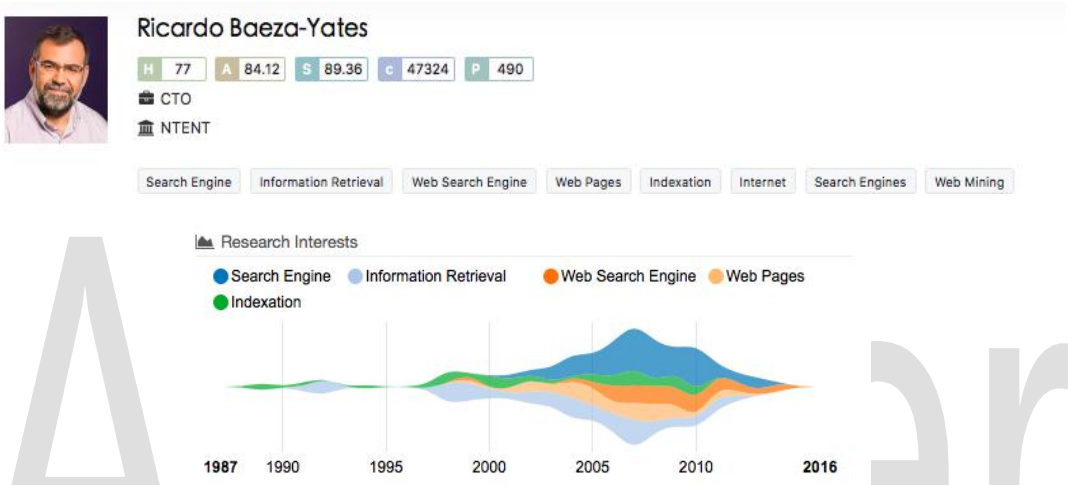


Susan Dumais 是美国计算机科学家，一直是微软搜索技术的重要贡献者。在加入微软之前，她曾在 Bellcore 实验室（现为 Telcordia Technologies）和同事进行关于信息检索的词汇问题（vocabulary problem）的研究，他们发现文档的作者可能使用与搜索文档的人不同的词汇来描述同一个事物，为了避免这个问题，他们发明了潜在语义索引（Latent Semantic Indexing）。

2006 年，Susan Dumais 当选 ACM Fellow；2009 年，她获得了被誉为信息检索领域终身成就奖的 Gerard Salton Award；2014 年，因为在计算机科学领域的杰出贡献，获得了雅典娜讲师奖（Athena Lecturer Award）；2015 年，入选美国艺术与科学学院（American Academy of Arts and Science）会士。

Susan Dumais 的研究领域包括：信息检索、人机互动等。她的高引论文“*Indexing by latent semantic analysis*”，描述了一种用于自动索引和检索的新办法。该办法是利用术语与文档（“语义结构”）的关联中的隐式高阶结构，以便基于查询中找到的术语改进相关文档的检测。

- Ricardo Baeza-Yates

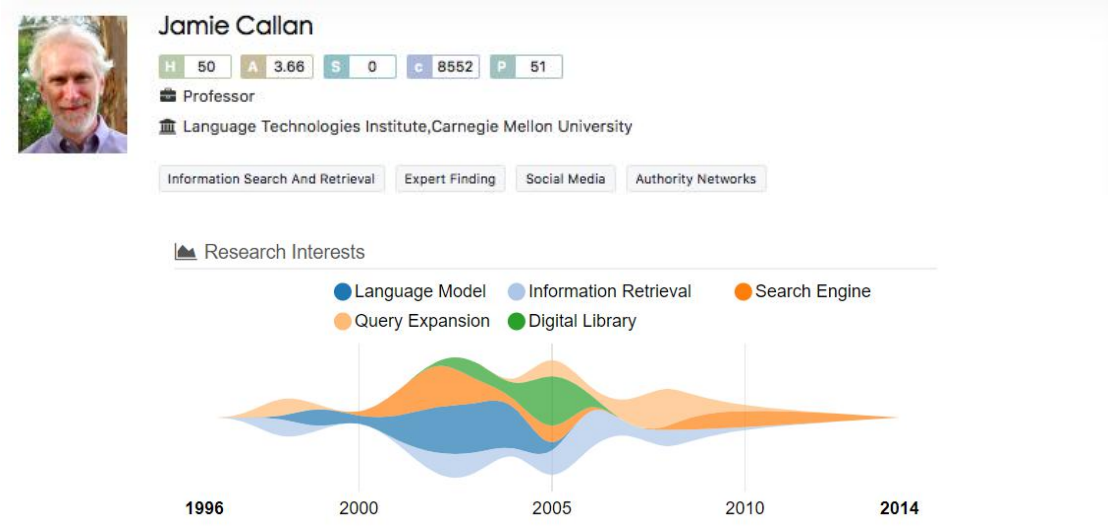


Ricardo Baeza-Yates 现任南加州语义搜索公司 NTENT 的首席技术官，东北大学（Northeastern University）硅谷校区的兼职教授。

2003 年，Ricardo Baeza-Yates 被选为智利科学院（Chilean Academy of Sciences）的成员；2009 年，被选为 ACM Fellow；2011 年，当选为 IEEE Fellow；2018 年被选为巴西科学院（Brazilian Academy of Sciences）成员。

Ricardo Baeza-Yates 的研究领域包括：算法和数据结构、信息检索、网络搜索和挖掘等。他的高引论文“*Data Portraits and Intermediary Topics: Encouraging Exploration of Politically Diverse Profiles*”提到，由于认知偏差，人们倾向于与志同道合的人进行互动，并且只阅读令人愉快的信息。许多努力让人们与那些思维方式不同的人联系起来并不顺利，因此他们尝试提出了一个混合推荐器算法和基于可视化的用户界面的平台来进行推荐。

- Jamie Callan

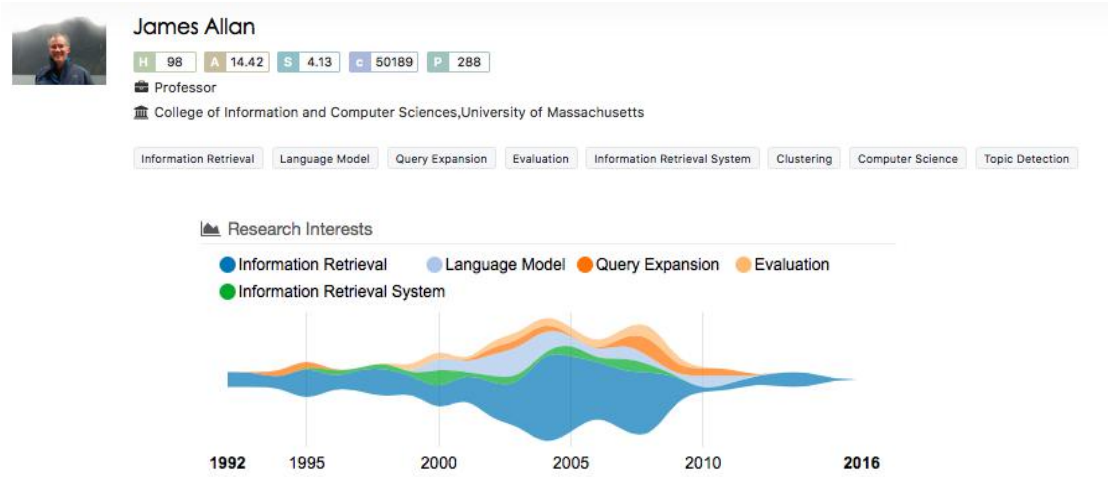


Jamie Callan 是卡内基·梅隆大学（Carnegie Mellon University）计算机系教授，之前曾在马赛诸塞大学（University of Massachusetts Amherst）阿默斯特分校担任助理教授。

Jamie Callan 是 ACM's Transactions on Information Systems（TOIS）期刊的主编，还担任过 ACM SIGIR 的前任财务主管（1999-2003）和主席（2003-2007）。

Jamie Callan 的研究领域包括：信息检索和分析、搜索引擎架构、信息过滤以及文本分析等。他的高引论文“*Distributed information retrieval*”，提出了一种分布式信息检索的多数据库模型，其中假设人们可以访问许多可搜索的文本数据库。在这样的环境中，全文信息检索包括发现数据库内容，按照预期满足查询的能力对数据库进行排序，搜索少量数据库，以及合并不同数据库返回的结果。

- James Allan

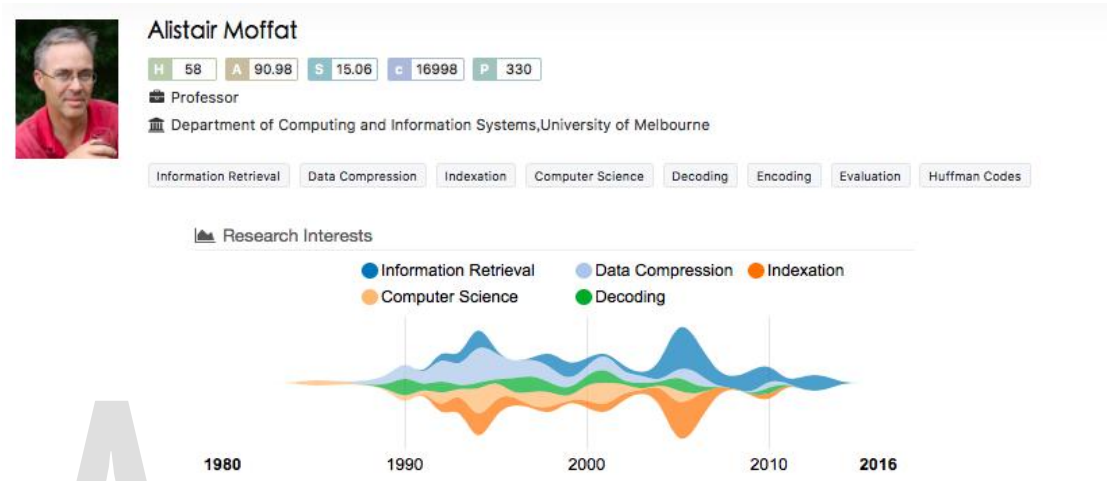


James Allan 是马赛诸塞大学阿默斯特分校（University of Massachusetts Amherst）计算机系教授，担任该学院主席，并且是智能中心的联合主任。

James Allan 曾担任 ACM's Transactions on Information Systems（TOIS）期刊和 Elsevier's Information Processing and Management（IPM）期刊的副主编。他目前是（Foundation and Trends in Information Retrieval）的编辑委员会成员，2018 年当选 CRA 董事会成员。

James Allan 研究的领域包括：信息检索等。他的高引论文“*On-line New Event Detection and Tracking*”，定义并描述了广播新闻中新事件检测和事件跟踪的相关问题。

- Alistair Moffat

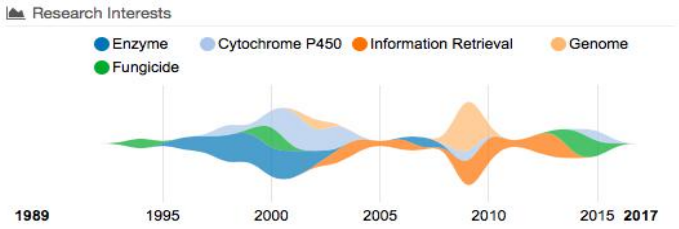


Alistair Moffat 是墨尔本大学（University of Melbourne）计算机系教授，担任计算机科学与软件工程系主任（2007-2011），墨尔本工程学院副院长（2007-2009）。

Alistair 担任过一系列会议的主席，也担任过期刊的副主编，包括 Journal of Information Retrieval 和 ACM Transactions on Information Systems（TOIS）等。

Alistair Moffat 研究的领域包括：文本和索引、信息检索、网络搜索等。他的高引论文：“*Self-Indexing Inverted Files for Fast Text Retrieval*”提到，大文本数据库的查询处理成本主要取决于检索和扫描每个查询项的反向列表的需要。通过使用压缩可以大大减少反转列表的检索时间，但是会增加所需的 CPU 时间。因此他们认为可以通过在每个压缩的反向列表中包含内部索引。

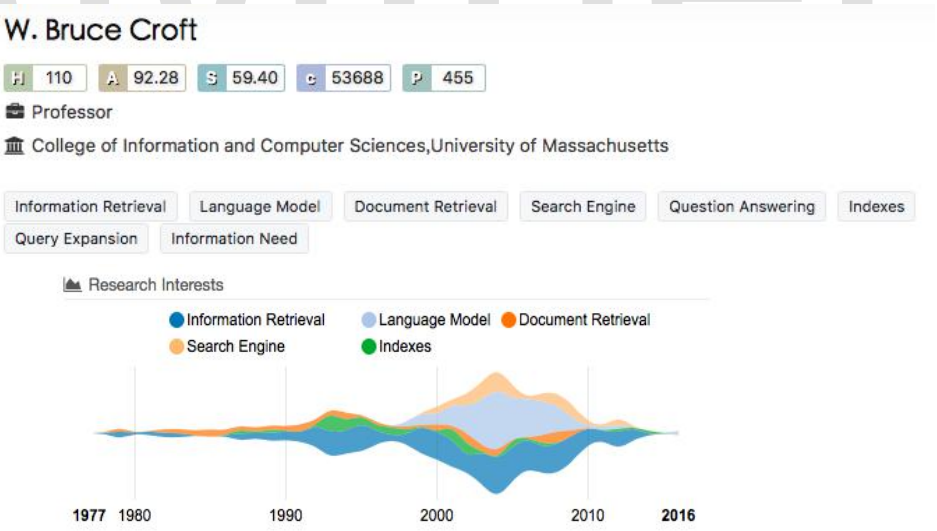
- Diane Kelly



Diane Kelly 是田纳西大学（University of Tennessee）计算机系教授，在此之前她曾供职于北卡罗来纳大学教堂山分校并担任教授。2004 年取得罗斯格大学的信息与图书馆学的博士学位；1996 年获得阿拉巴马大学（The University of Alabama）的心理学和英语学士学位。她曾获得 2014 年 ASIST 研究奖（每年只评一项）、2013 年英国计算机协会的 IRSG Karen Spärck Jones 奖、2009 年亚洲/汤森路透杰出信息科学教师奖和 2007 年 SILS 杰出教师奖。她是国际计算机学会信息检索分会（ACM SIGIR）的现任主席，同时是 ACM Transactions on Information Systems 副主编，并在包括 Information Processing & Management 和 Information Retrieval Journal 在内的多家核心期刊担任编委。

Diane Kelly 研究的领域包括：人机交互、信息搜索行为、交互式信息搜索等，她高引论文“*Methods for evaluating interactive information retrieval systems with users*”，提供了与用户评估交互式信息检索系统的概述和说明，并讨论了一些特定于交互式信息检索研究的伦理问题。

- W. Bruce Croft



W. Bruce Croft 是马萨诸塞大学（University of Massachusetts）阿默斯特分校的一位杰出的计算机科学教授，自 2015 年以来，一直担任信息与计算机科学院的院长。

W. Bruce Croft 是 ACM Fellow，CIIR（Center for Intelligent Information Retrieval）的创始人，并于 1995 年至 2002 年担任 ACM Transactions on Information Systems（TOIS）的主编。

1991 年成立的 CIIR 是 W. Bruce Croft 和他的学生，以及 90 多个行业和政府合作伙伴一起研究的技术项目，共发表了 900 多篇论文。Bruce Croft 为信息检索领域做出了重大贡献，在聚类、段落检索、句子检索和分布式搜索等方面都有开创性的工作成果。

W. Bruce Croft 的高引论文“*A language modeling approach to information retrieval*”，对文档索引和文档检索进行了广泛的研究。两类模型的整合一直是几位研究人员的目标，但这是一个非常棘手的问题，其中有很多原因是缺乏足够的索引模型。因此提出了一种基于概率语言建模的检索方法，单独估算每个文档的模型，用非参数化的建模方式将文档索引和文档检索集成到单个模型中。

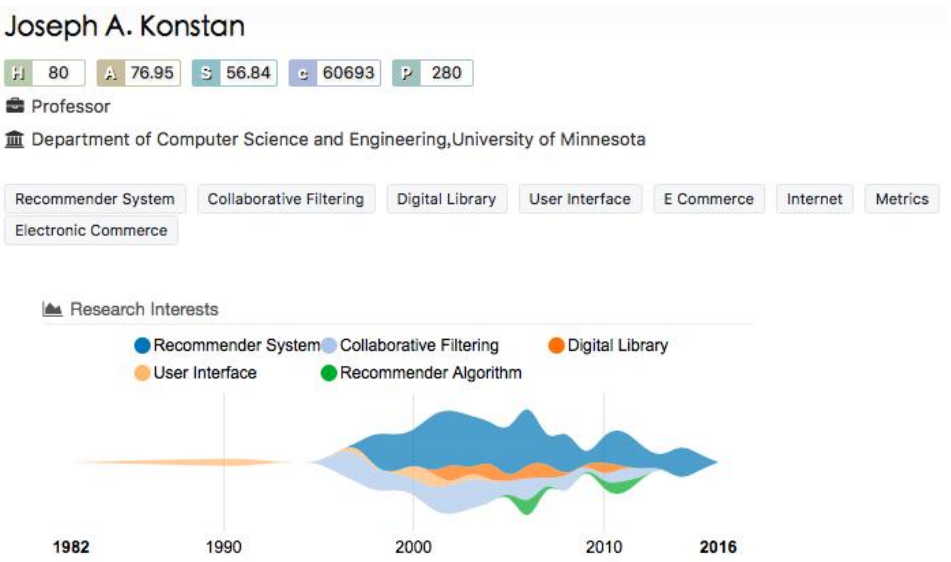
- Bernard J. Jansen

Bernard J. Jansen 是宾夕法尼亚州立大学（The Pennsylvania State University）信息科学与技术学院的教授，他毕业于西点军校，并拥有德克萨斯 A&M 大学（Texas A&M University，TAMU）的博士学位。

Bernard J. Jansen 不仅是 Information Processing & Management 期刊的主编，也是七个国际期刊的编辑委员会成员。他获得了多项奖项和荣誉，包括 ACM Research Award 等。

Bernard J. Jansen 的研究领域包括：网络搜索研究、搜索引擎研究和信息检索和检索研究等。他的高引论文“*Twitter Power: Tweets as Electronic Word of Mouth*”，分析了超过 150000 个包含品牌评论、情绪和意见的帖子。调查这些帖子的整体结构、表达方式以及积极或消极情绪的变化。将这些情绪分类的自动化方法与手动编码进行了比较。结果表明，19%的帖子都提到了一个品牌，在这些帖子里，超过 50%为正面，33%对公司或产品持批评态度。

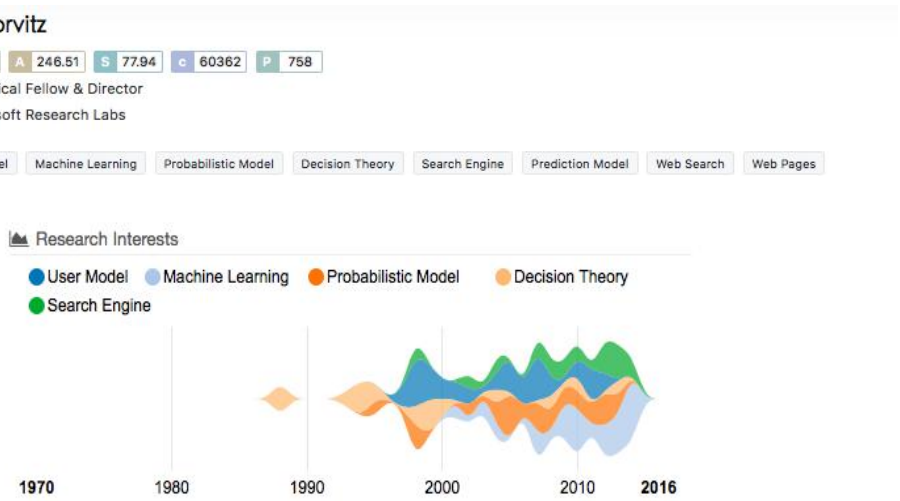
- Joseph A. Konstan



Joseph A. Konstan 是明尼苏达大学（University of Minnesota）计算机科学与工程学院的教授，他以协同过滤推荐人的工作以及他在线艾滋病预防方面的工作而闻名。

Joseph A. Konstan 的研究领域包括：社交计算、协作信息过滤，人机交互等。他的论文“*Item-based collaborative filtering recommendation algorithms*”提到，推荐系统将知识发现技术应用于在实时交互期间对信息、产品或服务进行个性化推荐的问题，尤其是基于 k 近邻协同过滤的系统。近年来，可用信息量和网站访问者数量的巨大增长为推荐系统带来了一些关键挑战：如何提供高质量的建议，为数百万用户和项目每秒执行许多建议，并在面对数据稀疏性时实现高覆盖率。

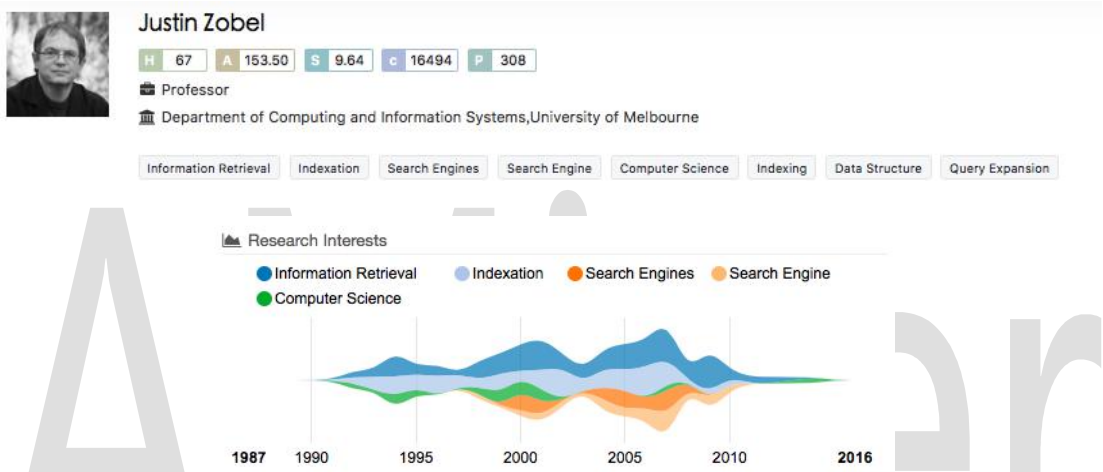
- Eric Horvitz



Eric Horvitz 是美国计算机科学家，微软技术研究员，在 2007 年-2009 年期间担任 AAAI 主席，在 2014 年被选为 ACM Fellow。他还获得过 2015 年的 AAAI Feigenbaum Prize 和 ACM-AAAI Allen Newell Award。

Eric Horvitz 的研究领域包括：开发感知、学习和推理系统的理论与实践，在机器学习和推理、信息检索、人机交互等领域作出了较大贡献。他的高引论文“*A Bayesian Approach to Filtering Junk E-Mail*”，提到为了解决互联网上日益严重的垃圾电子邮件问题，他们研究了自动构建过滤器的方法，即利用概率学习方法结合不同的错误分类成本的概念产生了特别适合于该任务的细微差别的过滤器。

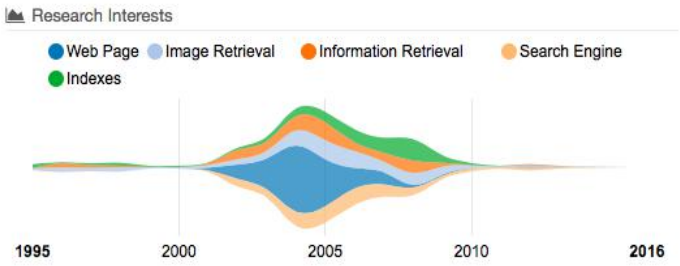
- Justin Zobel



Justin Zobel 是墨尔本大学（University of Melbourne）计算机与信息系统学院的教授，曾在 RMIT 和 NICTA 工作。2008 年，Justin Zobel 担任了澳大利亚部门和计算机科学学院 CORE 协会主席。

Justin Zobel 的研究领域包括：搜索技术、算法和生物信息学等。他的高引论文“*Inverted files for text search engines*”提到，在过去十年中，文本搜索引擎的基础技术已经取得了巨大进步，但许多具体技术并为广为人知，或者描述已经过时。因此 Justin Zobel 着重介绍了该领域的关键技术，描述核心实现以及如何通过一系列扩展来增强核心。

- 马维英

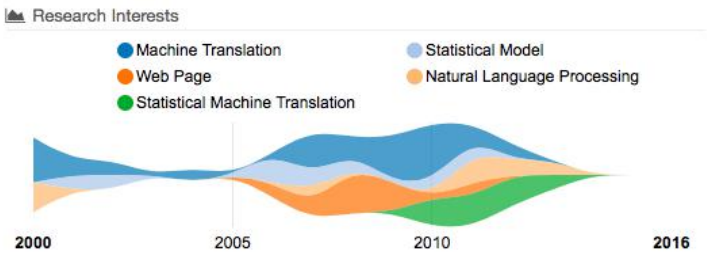


马维英是字节跳动的副总裁，领导人工智能（AI）实验室。他带领着团队进行基础研究，开发机器学习、自然语言计算、视觉计算、搜索、知识工程和人机交互等领域的新技术。

马维英在国际会议和期刊上发表了 270 多篇论文，获得了 160 多项专利，是 IEEE 的研究员和 ACM 的杰出科学家，曾经担任 2008 年万维网国际会议（WWW）的项目联合主席，以及 2011 年信息检索特别兴趣小组（SIGIR）的联合主席。

马维英的高引论文“*Texture Features for Browsing and Retrieval of Image Data*”，提到了基于图像内容的检索正在成为一个重要的研究领域，应用于数字图书馆和多媒体数据库。论文的重点是图像处理方面，特别是使用文本信息来浏览和检索大图像数据，建议使用 Gabor 小波特征进行文本分析，并提供全面的实验评估。

- 周明



周明是微软亚洲研究院副院长，ACL 候选主席，中国中文信息学会常务会理事。他带领团队进行了微软输入法、英库词典（必应词典）、中英翻译、微软中国文化系列（微软对联、微软字谜、微软绝句）等重要产品和项目的研发。

周明发表了 120 余篇重要会议和期刊论文，拥有国际发明专利 40 余项。他的高引论文“*基于层次结构的多策略中文微博情感分析和特征抽取*”提到，随着 Web2.0 时代的兴起，

与微博相关的研究得到了学术界和工业界的广泛关注，文中使用新浪 API 获取数据，对中文微博消息展开了情感分析研究。

3.2.2 中青年学者

- 李航

李航是字节跳动 AI 实验室主任，1990-2001 年在 NEC 公司的研究实验室工作；2001-2012 年在微软亚洲研究院工作；2012-2017 年在华为技术公司工作。

李航的研究领域包括：自然语言处理、信息检索、机器学习和数据挖掘等。他的高引论文“*AdaRank：a boosting algorithm for information retrieval*”，提出了解决学习文档检索排名的新方法，即在任务中，使用一些训练数据自动创建模型。

- 周涛

周涛，电子科技大学互联网科学中心主任，长期从事复杂性科学和大数据挖掘算法和应用研究，在 Physics Reports 和 PNAS 等国际顶尖 SCI 期刊发表论文 200 余篇，多次获得 2014 和 2015 年连续两年入选 Elsevier 最具国际影响力中国科学家名单。

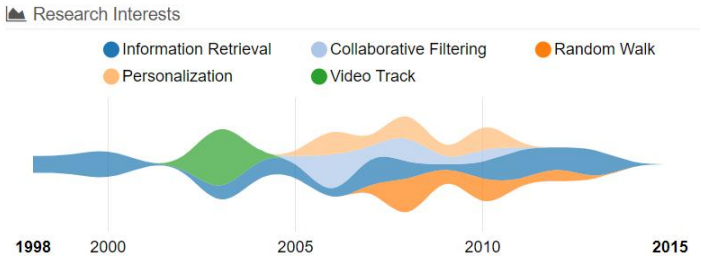
周涛的高引论文“*个性化推荐系统的研究进展*”提到，互联网技术的迅猛发展把我们带进了信息爆炸的时代。个性化推荐系统通过建立用户与信息产品之间的二元关系，利用已有的选择过程或相似性关系挖掘每个用户潜在感兴趣的对象，进而进行个性化推荐，其本质就是信息过滤。

- Emine Yilmaz

Emine Yilmaz 是伦敦大学学院（University College London）计算机科学系教授。她担任过 ACM SIGIR 2018 的 PC 主席，2017 年的 ECIR 的博士联合主席。

Emine Yilmaz 的研究领域包括：信息检索、数据挖掘以及机器学习等。她的高引论文“*A new rank correlation coefficient for information retrieval*”提到，在信息检索领域，人们经常面临计算两个排名列表之间的相关性的问题。在原有的肯德尔的统计数据的基础上，Emine Yilmaz 提出了秩相关系数，即 AP 相关性（Tap），它基于平均精度并具有概率解释。

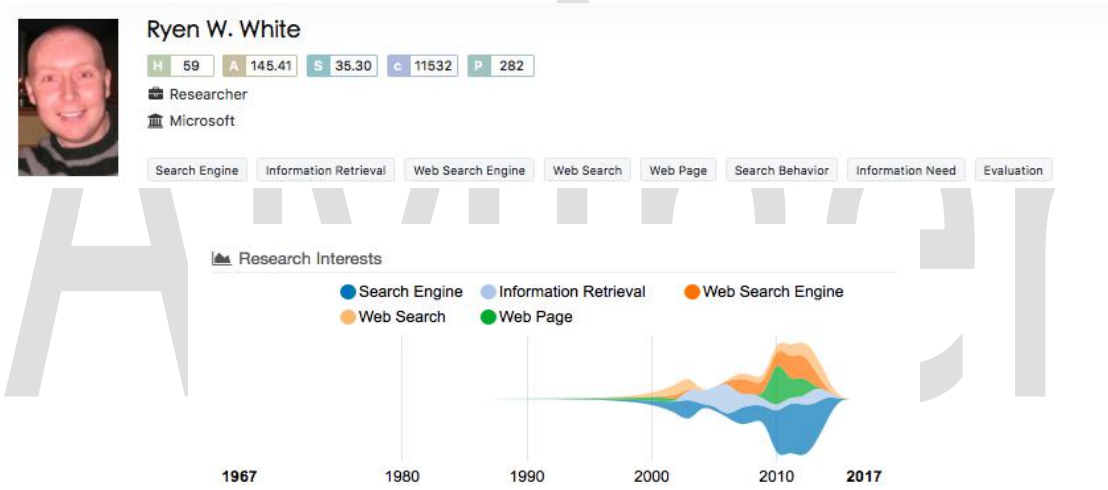
- Arjen de Vries



Arjen de Vries 是 Radboud University 的教授，担任过 2012 年 ECIR 的 PC 主席，2007 年 SIGIR 的主席。

Arjen de Vries 的研究领域包括：推荐系统、协同过滤等。他的高引论文“*Unifying User-based and Item-based Collaborative Filtering Approaches by Similarity Fusion*”提到，由于评级数据固有的稀疏性，无法获得来自类似用户或类似项目的大量评级，导致预测质量差。文章在生成概率框架中重新制定了基于记忆的协同过滤问题，将个人用户项目评级视为缺失评级的预测因子。

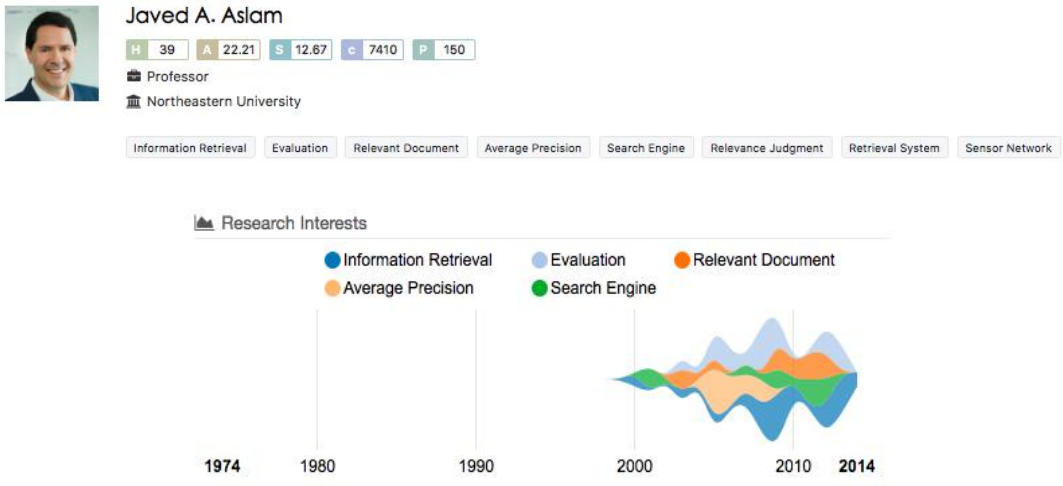
- Ryen White



Ryen White 是 Microsoft Research AI 的研究员，他还是 ACM Transction on Information Systems（TOIS）和 ACM Transaction on the Web 的编辑。Ryen White 获得过三次 SIGIR Best Paper Award，分别是 2007 年、2010 年和 2013 年。

Ryen White 的研究领域包括搜索引擎、网页搜索等，他的高引论文“*Predicting users interest from contextual information*”提出了系统研究五种不同的情境信息来源对用户兴趣建模的有效性。使用的五个上下文信息来源是：社交、历史、任务、集合和用户交互，根据他们预测用户未来兴趣的有效程度来评估这些来源的效用和它们之间的重叠。

- Javed Aslam



Javed Aslam 是东北大学（Northeastern University）的教授，在加入东北大学之前，曾在达特茅斯学院担任计算机科学系的助理教授。Javed Aslam 还曾担任过 2009 年的 ACM SIGIR 的联合主席。

Javed Aslam 的研究领域包括：信息检索、机器学习、算法的设计和分析等。他的高引论文“*Models for Metasearch*”对元搜索问题做出了三点贡献，描述和研究了基于最优民主投票程序 Borda Count 的元搜索模型；描述和研究了基于贝叶斯推理的元搜索模型；探索了用于获得元搜索算法性能的上界。

3.3 论文介绍

3.3.1 近年 ACM SIGIR 获奖论文

本节列举了 2016 至 2018 年度 ACM SIGIR 获奖论文，论文题目、作者、摘要信息以及下载链接如下：

★ SIGIR 2018 获奖论文

- **Best paper award**

论文题目：***Should I Follow the Crowd? A Probabilistic Analysis of the Effectiveness of Popularity in Recommender Systems***

论文作者：Rocío Cañamares, Pablo Castells

论文摘要：The use of IR methodology in the evaluation of recommender systems has become common practice in recent years. IR metrics have been found however to be strongly biased

towards rewarding algorithms that recommend popular items –the same bias that state of the art recommendation algorithms display. Recent research has confirmed and measured such biases, and proposed methods to avoid them. The fundamental question remains open though whether popularity is really a bias we should avoid or not; whether it could be a useful and reliable signal in recommendation, or it may be unfairly rewarded by the experimental biases. We address this question at a formal level by identifying and modeling the conditions that can determine the answer, in terms of dependencies between key random variables, involving item rating, discovery and relevance. We find conditions that guarantee popularity to be effective or quite the opposite, and for the measured metric values to reflect a true effectiveness, or qualitatively deviate from it. We exemplify and confirm the theoretical findings with empirical results. We build a crowdsourced dataset devoid of the usual biases displayed by common publicly available data, in which we illustrate contradictions between the accuracy that would be measured in a common biased offline experimental setting, and the actual accuracy that can be measured with unbiased observations.

论文地址：https://www.aminer.cn/archive/should-i-follow-the-crowd-a-probabilistic-analysis-of-the-effectiveness-of-popularity-in-recommender-systems/5b67b46f17c44aac1c8632a4

- **Test of time award**

论文题目：***Improving web search ranking by incorporating user behavior information***

论文作者：Eugene Agichtein, Eric Brill, Susan Dumais

论文摘要：We show that incorporating user behavior data can significantly improve ordering of top results in real web search setting. We examine alternatives for incorporating feedback into the ranking process and explore the contributions of user feedback compared to other common web search features. We report results of a large scale evaluation over 3,000 queries and 12 million user interactions with a popular web search engine. We show that incorporating implicit feedback can augment other features, improving the accuracy of a competitive web search ranking algorithms by as much as 31% relative to the original performance.

论文地址：https://www.aminer.cn/archive/improving-web-search-ranking-by-incorporating-user-behavior-information/53e9b042b7602d9703aa115b

★ SIGIR 2017 获奖论文

- **Best paper award**

论文题目：***BitFunnel: Revisiting Signatures for Search***

论文作者：Bob Goodwin, Michael Hopcroft, Dan Luu, Alex Clemmer, Mihaela Curmei, Sameh Elnikety, Yuxiong He

论文摘要：Since the mid-90s there has been a widely-held belief that signature files are inferior to inverted files for text indexing. In recent years the Bing search engine has developed and deployed an index based on bit-sliced signatures. This index, known as BitFunnel, replaced an existing production system based on an inverted index. The driving factor behind the shift away from the inverted index was operational cost savings. This paper describes algorithmic innovations and changes in the cloud computing landscape that led us to reconsider and eventually field a technology that was once considered unusable. The BitFunnel algorithm directly addresses four fundamental limitations in bit-sliced block signatures. At the same time, our mapping of the algorithm onto a cluster offers opportunities to avoid other costs associated with signatures. We show these innovations yield a significant efficiency gain versus classic bit-sliced signatures and then compare BitFunnel with Partitioned Elias-Fano Indexes, MG4J, and Lucene.

论文地址：

https://www.aminer.cn/archive/bitfunnel-revisiting-signatures-for-search/59a030a9b161e8ad1a7b6f13

- **Test of time award**

论文题目：***Personalizing search via automated analysis of interests and activities***

论文作者：Jaime Teevan, Susan T. Dumais, Eric Horvitz

论文摘要：We formulate and study search algorithms that consider a user's prior interactions with a wide variety of content to personalize that user's current Web search. Rather than relying on the unrealistic assumption that people will precisely specify their intent when searching, we pursue techniques that leverage implicit information about the user's interests. This information is used to re-rank Web search results within a relevance feedback framework. We explore rich models of user interests, built from both search-related information, such as previously issued queries and previously visited Web pages, and other information about the user such as documents and email the user has read and created. Our research suggests that rich representations of the user and the corpus are important for personalization, but that it is possible to approximate these representations and provide efficient client-side algorithms for personalizing search. We show that such personalization algorithms can significantly improve on current Web search.

论文地址：

https://www.aminer.cn/archive/personalizing-search-via-automated-analysis-of-inte-rests-and-activities/53e9a53fb7602d9702e60c49

★ SIGIR 2016 获奖论文

- **Best paper award**

论文题目：***Understanding Information Need: an FMRI Study***

论文作者：Yashar Moshfeghi, Peter Triantafillou, Frank E. Pollick

论文摘要：Information need refers to a complex concept: at the very initial state of the phenomenon (i.e. at a visceral level), even the searcher may not be aware of its existence. This renders the measuring of this concept (using traditional behaviour studies) nearly impossible. In this paper, we investigate the connection between an information need and brain activity. Using functional Magnetic Resonance Imaging (fMRI), we measured the brain activity of twenty-four participants while they performed a Question Answering (Q/A) Task, where the questions were carefully selected and developed from TREC-8 and TREC 2001 Q/A Track. The results of this experiment revealed a distributed network of brain regions commonly associated with activities related to information need and retrieval and differing brain activity in processing scenarios when participants knew the answer to a given question and when they did not and needed to search.

论文地址：

https://www.aminer.cn/archive/understanding-information-need-an-fmri-study/57d063b4ac4436735428dbe2

- **Test of time award**

论文题目：***Accurately interpreting Click through data as implicit feedback***

论文作者：Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, Geri Ga

论文摘要：This paper examines the reliability of implicit feedback generated from clickthrough data in WWW search. Analyzing the users' decision process using eyetracking and comparing implicit feedback against manual relevance judgments, we conclude that clicks are informative but biased. While this makes the interpretation of clicks as absolute relevance judgments difficult, we show that relative preferences derived from clicks are reasonably accurate on average.

论文地址：

https://www.aminer.cn/archive/accurately-interpreting-clickthrough-data-as-implicit-feedback/53e9aacab7602d9703447ca7

3.3.2 近五年 ACM SIGIR 高引论文

本节列举了 ACM SIGIR 会议近五年高引论文，论文题目、作者、摘要信息以及下载链接如下：

- 题目：***Image-Based Recommendations on Styles and Substitutes***（引用数：572）

作者：Julian J. McAuley, Christopher Targett, Qinfeng Shi, Anton van den Hengel

摘要：Humans inevitably develop a sense of the relationships between objects, some of which are based on their appearance. Some pairs of objects might be seen as being alternatives to each other (such as two pairs of jeans), while others may be seen as being complementary (such as a pair of jeans and a matching shirt). This information guides many of the choices that people make, from buying clothes to their interactions with each other. We seek here to model this human sense of

the relationships between objects based on their appearance. Our approach is not based on fine-grained modeling of user annotations but rather on capturing the largest dataset possible and developing a scalable method for uncovering human notions of the visual relationships within. We cast this as a network inference problem defined on graphs of related images, and provide a large-scale dataset for the training and evaluation of the same. The system we develop is capable of recommending which clothes and accessories will go well together (and which will not), amongst a host of other applications.

论文地址：https://www.aminer.cn/archive/573696c56e3b12023e5c58be

- 题目：***Learning to Rank Short Text Pairs with Convolutional Deep Neural Networks***（引用数：419）

作者：Aliaksei Severyn, Alessandro Moschitti

摘要：Learning a similarity function between pairs of objects is at the core of learning to rank approaches. In information retrieval tasks we typically deal with query-document pairs, in question answering -- question-answer pairs. However, before learning can take place, such pairs needs to be mapped from the original space of symbolic words into some feature space encoding various aspects of their relatedness, e.g. lexical, syntactic and semantic. Feature engineering is often a laborious task and may require external knowledge sources that are not always available or difficult to obtain. Recently, deep learning approaches have gained a lot of attention from the research community and industry for their ability to automatically learn optimal feature representation for a given task, while claiming state-of-the-art performance in many tasks in computer vision, speech recognition and natural language processing. In this paper, we present a convolutional neural network architecture for reranking pairs of short texts, where we learn the optimal representation of text pairs and a similarity function to relate them in a supervised way from the available training data. Our network takes only words in the input, thus requiring minimal preprocessing. In particular, we consider the task of reranking short text pairs where elements of the pair are sentences. We test our deep learning system on two popular retrieval tasks from TREC: Question Answering and Microblog Retrieval. Our model demonstrates strong performance on the first task beating previous state-of-the-art systems by about 3\% absolute points in both MAP and MRR and shows comparable results on tweet reranking, while enjoying the benefits of no manual feature engineering and no additional syntactic parsers.

论文地址：https://www.aminer.cn/archive/573696c46e3b12023e5c5745

- 题目：***Attentive Collaborative Filtering: Multimedia Recommendation with Item- and Component-Level Attention***（引用数：167）

作者：Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Weiwei Liu, Tat-Seng Chua

摘要：Multimedia content is dominating today's Web information. The nature of multimedia user-item interactions is 1/0 binary implicit feedback (e.g., photo likes, video views, song downloads, etc.), which can be collected at a larger scale with a much lower cost than explicit feedback (e.g., product ratings). However, the majority of existing collaborative filtering (CF) systems are not well-designed for multimedia recommendation, since they ignore the implicitness in users' interactions with multimedia content. We argue that, in multimedia recommendation, there exists item- and component-level implicitness which blurs the underlying users' preferences. The item-level implicitness means that users' preferences on items (e.g. photos, videos, songs, etc.) are unknown, while the component-level implicitness means that inside each item users' preferences on different components (e.g. regions in an image, frames of a video, etc.) are unknown. For example, a 'view" on a video does not provide any specific information about how the user likes the video (i.e.item-level) and which parts of the video the user is interested in (i.e.component-level). In this paper, we introduce a novel attention mechanism in CF to address the challenging item- and component-level implicit feedback in multimedia recommendation, dubbed Attentive Collaborative Filtering (ACF). Specifically, our attention model is a neural network that consists of two attention modules: the component-level attention module, starting from any content feature extraction network (e.g. CNN for images/videos), which learns to select informative components of multimedia items, and the item-level attention module, which learns to score the item preferences. ACF can be seamlessly incorporated into classic CF models with implicit feedback, such as BPR and SVD++, and efficiently trained using SGD. Through extensive experiments on two real-world multimedia Web services: Vine and Pinterest, we show that ACF significantly outperforms state-of-the-art CF methods.

论文地址：https://www.aminer.cn/archive/599e64305d763cf99f73a2d0

- 题目：***Rank-GeoFM: A Ranking based Geographical Factorization Method for Point of Interest Recommendation***（引用数：160）

作者：Xutao Li, Gao Cong, Xiao-Li Li, Tuan-Anh Nguyen Pham, Shonali Krishnaswamy

摘要：With the rapid growth of location-based social networks, Point of Interest (POI) recommendation has become an important research problem. However, the scarcity of the check-in data, a type of implicit feedback data, poses a severe challenge for existing POI recommendation methods. Moreover, different types of context information about POIs are available and how to leverage them becomes another challenge. In this paper, we propose a ranking based geographical factorization method, called Rank-GeoFM, for POI recommendation, which addresses the two challenges. In the proposed model, we consider that the check-in frequency characterizes users' visiting preference and learn the factorization by ranking the POIs correctly. In our model, POIs both with and without check-ins will contribute to learning the ranking and thus the data sparsity problem can be alleviated. In addition, our model can easily incorporate different types of context information, such as the geographical influence and

temporal influence. We propose a stochastic gradient descent based algorithm to learn the factorization. Experiments on publicly available datasets under both user-POI setting and user-time-POI setting have been conducted to test the effectiveness of the proposed method. Experimental results under both settings show that the proposed method outperforms the state-of-the-art methods significantly in terms of recommendation accuracy.

论文地址：https://www.aminer.cn/archive/573696c56e3b12023e5c59f6

- 题目：***Monolingual and Cross-Lingual Information Retrieval Models Based on (Bilingual) Word Embeddings***（引用数：150）

作者：Ivan Vulić, Marie-Francine Moens

摘 要： We propose a new unified framework for monolingual (MoIR) and cross-lingual information retrieval (CLIR) which relies on the induction of dense real-valued word vectors known as word embeddings (WE) from comparable data. To this end, we make several important contributions: (1) We present a novel word representation learning model called Bilingual Word Embeddings Skip-Gram (BWESG) which is the first model able to learn bilingual word embeddings solely on the basis of document-aligned comparable data; (2) We demonstrate a simple yet effective approach to building document embeddings from single word embeddings by utilizing models from compositional distributional semantics. BWESG induces a shared cross-lingual embedding vector space in which both words, queries, and documents may be presented as dense real-valued vectors; (3) We build novel ad-hoc MoIR and CLIR models which rely on the induced word and document embeddings and the shared cross-lingual embedding space; (4) Experiments for English and Dutch MoIR, as well as for English-to-Dutch and Dutch-to-English CLIR using benchmarking CLEF 2001-2003 collections and queries demonstrate the utility of our WE-based MoIR and CLIR models. The best results on the CLEF collections are obtained by the combination of the WE-based approach and a unigram language model. We also report on significant improvements in ad-hoc IR tasks of our WE-based framework over the state-of-the-art framework for learning text representations from comparable data based on latent Dirichlet allocation (LDA).

论文地址：https://www.aminer.cn/archive/573696c56e3b12023e5c5f0c

4 产业应用篇

在不同的领域中，信息的关联方式都存在一定的差异，因此本文选取了各个领域具有代表性的例子来进行分析。

4.1 典型技术应用产品

- **百度搜索**

2000 年 1 月李彦宏、徐勇两人在北京中关村创立了百度搜索。百度搜索的核心技术是“超链接分析”，它通过分析链接网站的多少来评价被链接的网站质量，这保证了用户在使用搜索引擎的时候，越受用户欢迎的内容排名越靠前。

百度搜索提供网页搜索、MP3 搜索、图片搜索、新闻搜索、百度百科等主要产品和服务，同时还提供更加细分的搜索服务，如地图搜索，邮编搜索等服务。

目前百度搜索拥有全球最大的中文网页，是全球第二大搜索引擎。百度搜索在中国搜索引擎市场占 76.05%的市场份额，截止 2018 年 5 月，它的市值上升至 990 亿美元。

- **今日头条**

2012 年 3 月张一鸣创建了今日头条，并于同年 8 月发布了第一个版本。2005 年，今日头条获得了年度最具影响力 APP 奖。

今日头条是北京字节跳动科技有限公司开发的，它是基于数据挖掘的推荐搜索引擎产品，内容大部分是从网上抓取的。因此，头条的特色是基于个性化推荐引擎技术，即根据每个用户的兴趣、位置等多个维度挖掘出其兴趣，再进行推荐，推荐内容包括新闻、音乐、电影、游戏、购物等资讯。

2016 年 9 月 20 日，今日头条投资 10 亿元用以补贴短视频创作，孵化 UGC 短时频平台“火山小视频”，迅速抢滩视频产业。目前，作为聚合类新闻客户端的今日头条的用户体量已经达到了行业第二。

- **豆瓣**

豆瓣是杨勃于 2005 年 3 月创立的社区网站。它除了提供书籍、电影、音乐等作品的信息外，还提供线下同城活动、小组话题讨论等多种服务功能。

以书影音起家的豆瓣，是 Web2.0 时代极具影响力的创新企业。在豆瓣上，用户们可以自由发表有关书影音的评论，也可以搜索别人的推荐，所有的内容、分类、筛选、排序等都是由用户生产和决定的。

2012 年，豆瓣的月度覆盖独立用户数已超过 1 亿，目前豆瓣、猫眼等网站承载了人们在影视文娱方面需求的重要功能。

4.2 垂直应用

- **法律搜索**

法律搜索能提供法律法规、案例等法律信息以及与之相关的时事新闻、商业资讯。

如 Westlaw International，它是全球最大的在线法律研究工具，用户们可以通过它迅速地存取案例、法令法规、表格、条约、商业资料和更多的资源。

- **健康搜索**

健康搜索可以为用户们提供疾病、保健、医生、医院等方面的信息。

如 iMedix，它是美国近几年崛起的一个关于健康主题的搜索引擎与博客社区，它整合了搜索和社交网络功能，改变了人们在线寻找健康信息的办法，用户被鼓励彼此帮助来共享健康体验。

- **问答式搜索**

面对具体任务的问答系统维护成本较高，大部分由手写规则构成，扩展能力较差。少部分较为先进的系统采用了检索式方案，它的本质是对用户问题进行分类再针对性回答。问答式搜索的核心是对用户的问句与已有的问句/答句库中的所有候选问句计算语义相似度并列序选出最相似的问句，再使用此问句的对应答句回答用户问句。

如 Quora，它是一个问答 SNS 网站，由 Facebook 前雇员查理·切沃和亚当·安捷罗在 2009 年创立的。网站集合了很多问题和答案，也容许用户协同编辑问题和答案。

4.3 产品推荐

产品推荐算法广泛应用于淘宝、京东、亚马逊等购物网站。在产品推荐中广泛使用的协同过滤算法（Collaborative Filtering），协同过滤推荐算法是诞生最早，并且较为著名的推荐算法。主要的功能是预测和推荐。算法通过对用户历史行为数据的挖掘发现用户的偏好，基于不同的偏好对用户进行群组划分并推荐品味相似的商品。协同过滤推荐算法分为两类，分别是基于用户的协同过滤算法（user-based collaborative filtering），和基于物品的协同过滤算法（item-based collaborative filtering），简单的说就是：人以类聚，物以群分[27]。因此这些网站可以根据用户以前的购物记录，推荐用户可能感兴趣的产品。

- 基于用户的协同过滤算法

基于用户的协同过滤算法是通过用户的历史行为数据发现用户对商品或内容的喜欢（如商品购买，收藏，内容评论或分享），并对这些喜好进行度量和打分。根据不同用户对相同商品或内容的态度和偏好程度计算用户之间的关系。在有相同喜好的用户间进行商品推荐。例如，小明和小红两个用户都购买了《三体》、《火星编年史》、《安德的游戏》三本图书，并且给出了 5 星的好评。那么小明和小红就属于同一类用户。可以将小明看过的图书《银河系漫游指南》也推荐给用户小红。

- 基于物品的协同过滤算法

基于物品的协同过滤算法与基于用户的协同过滤算法很像，将商品和用户互换。通过计算不同用户对不同物品的评分获得物品间的关系。基于物品间的关系对用户进行相似物品的推荐。这里的评分代表用户对商品的态度和偏好。简单来说就是如果小颖同时购买了商品笔记本和签字笔，那么说明笔记本和签字笔的相关度较高。当用户小彬也购买了笔记本时，可以推断他也有购买签字笔的需求。

4.4 音乐推荐

信息推荐也广泛运用于音乐推荐领域，即根据用户以前的听歌记录，推荐用户可能感兴趣的歌曲。例如网易云音乐，它主要根据每日获取到的听歌列表，优先推荐跟该歌曲相似的歌曲[28]。目前流行的音乐推荐算法主要分为基于内容的推荐算法、协同过滤推荐算法等。

- 基于内容的推荐算法

利用用户的听歌历史，获取用户的音乐偏好信息，推荐的是和用户之前听的相似度很高。例如：用户 A 喜欢歌曲 a（特征：华语、摇滚）；用户 B 喜欢歌曲 b（特征：欧美、爵士）。当一个新的歌曲 c 提取到的特征是（欧美、爵士），由歌曲 c 和用户 B 的高匹配度可以将歌曲 c 推荐给用户 B，如下图所示：

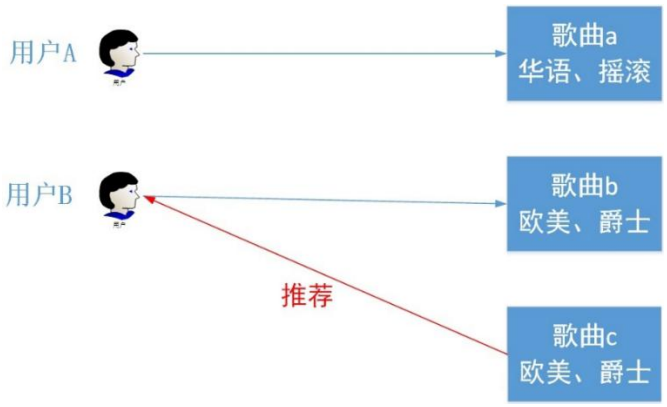


图 16 音乐推荐基于内容的推荐算法

- 协同过滤推荐算法

基于用户的协同过滤推荐：当用户 A 需要个性化音乐推荐时，系统帮他找到和他有相似兴趣的用户群集合 M，和该用户群喜欢的音乐集合 N；之后系统会预测目标用户 A 对音乐集合 N 的评分，然后对音乐集合 N 按评分由高到低的顺序推荐给目标用户 A。这就是基于用户的协同过滤。例如：用户 A 喜欢歌曲 a，歌曲 c，歌曲 d。目标用户 C 喜欢歌曲 a，歌曲 c。即用户 A、C 可能有相同音乐偏好。所以系统把歌曲 d 推荐给用户 C，如下图所示：

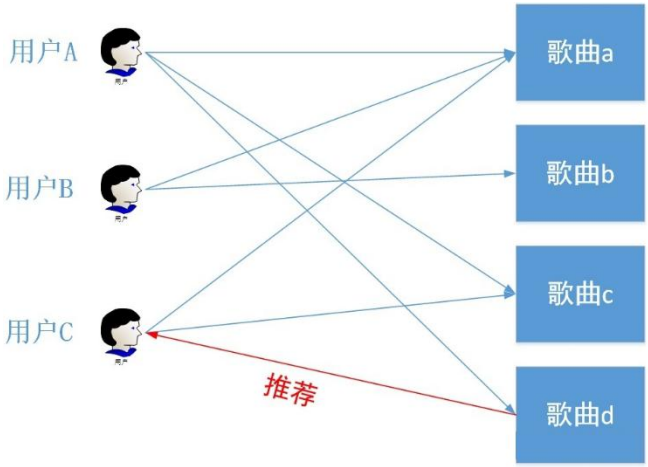


图 17 音乐推荐基于用户的协同过滤推荐

基于商品的协同过滤推荐：根据所有用户对音乐的偏好，计算系统中任意两个音乐间的相似度，然后根据用户的历史偏好物品列表，将列表中物品相似度较高的物品推荐给目标用户。例如：假设用户 A 喜欢歌曲 a，b，c;用户 B 喜欢 a，c，d。可以看出喜欢了歌曲 a 的用户都喜欢了歌曲 c，则分析得出歌曲 a 和 c 比较相似。系统发现用户 C 喜欢了歌曲 a,则可以把歌曲 a 的相似歌曲 c 推荐给用户 C，如下图所示：

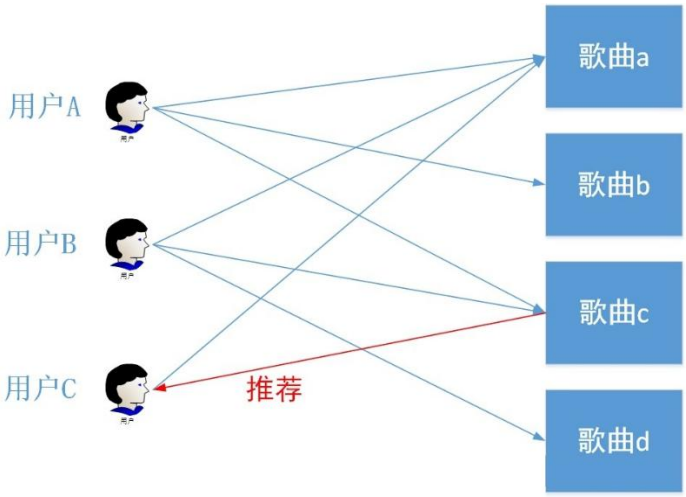


图 18 音乐推荐基于商品的协同过滤推荐

4.5 信息流推荐

信息流产品设计领域非常多，包括内容库、用户画像、短视频、搜索、信息流广告等。它的产品形态具有以下几个特点：海量信息，能源源不断地刷出新的、实时的内容；能在合适的场景下，为用户提供合适的内容；用户黏性强、使用时间长，利于广告曝光创造营收。例如今日头条，它就是基于数据挖掘的推荐引擎产品，能够实现 5 秒快速推广，锁定目标用户，10 秒更新用户模型，实现更加精确的广告投放。头条具有强大的推荐系统，可给定向人群投放定制素材。信息流非常大地拓宽了媒体拥有的广告位数量，同时避免对用户不必要的骚扰。同时，信息流以传统广告模式结合新媒体技术（大数据、人工智能、受众画像），通过优质媒体，主动向潜在用户提供易于接受营销信息，为广告主们提供全新的营销蓝海市场。

信息流推荐内容的方式包括人工运营、算法推荐以及两者的结合，人工和算法各有所长：人工运营更擅长新闻价值判断以及对热点的预测；而算法更适合运用于个性化匹配、冷门的长尾内容推荐上。大数据时代，每天更新的内容是海量的。而人工运营，往往局限于热点内容，就像是冰山的一角。冰山之下，是大量的长尾、冷门的内容，必须依赖机器算法做个性化推荐。目前信息流推荐的主流算法架构是召回和排序，召回算法最终决定了推荐效果的上界，而排序则保证了推荐结果的精准。所以从模型优化的角度来讲，只有保证召回和排序双管齐下，才能发挥推荐系统的最好效果。

在信息流领域，百度正在推行“搜索+信息流”战略，百度信息流是百度在 2016 年第二季度正式推出的一个个性化阅读应用。百度副总裁沈抖在接受媒体采访时表示，用户不仅仅可以在手机百度的首页上看到百度信息流，也可以在百度浏览器首页、百度移动端首页等其他地方看到百度信息流。在搜索的基础上，增加信息流产品，通过二者的双向互补循环，帮助用户不仅能更智能地搜到自己想要的信息，人工智能技术还能根据用户的需求，主动推荐信息到用户眼前，实现不搜即得。对于商家而言，信息流广告分发效率也更加精准，比展示广告更能传递品牌信息。百度搜索+推荐的模式带动下的信息流品台广告的价值也得到了业界广泛认可。沈抖表示，新移动时代，百度信息流将随时随地、无处不在的主动为用户提供全面的信息服务。而强大的人工智能技术和多赢的百度信息流生态，则保证了手机百度能给用户提供世界上最好的信息流产品。在今年的 KDD 研讨会上，沈抖发表了题为“Latest Practice in Newsfeed Recommendation Engine”的演讲，介绍了百度新闻推荐引擎、用户建模的主要技术、推荐算法等，让大家更深入的认识了百度信息流发展情况。

5 趋势篇

随着网络时代的来临，信息检索与推荐技术得到了迅猛的发展，未来信息检索与推荐将呈现出更加智能化、标准化、个性化的发展态势。

5.1 发展关键词回顾

随着网络技术的发展，信息资源也呈现出爆炸性增长，手工检索已经很难适应当今信息的发展速度，计算机检索应运而生。

AMiner 根据信息检索与推荐领域的相关论文绘制研究热点趋势图（如图 19），旨在基于历史的科研成果数据的基础上，对信息检索与推荐各个时间段的热度和发展趋势进行研究。图中，每个色彩分支表示一个关键词领域，其宽度表示关键词的研究热度，各关键词在每一年份（纵轴）的位置是按照这一时间点上所有关键词的热度高低进行排序。

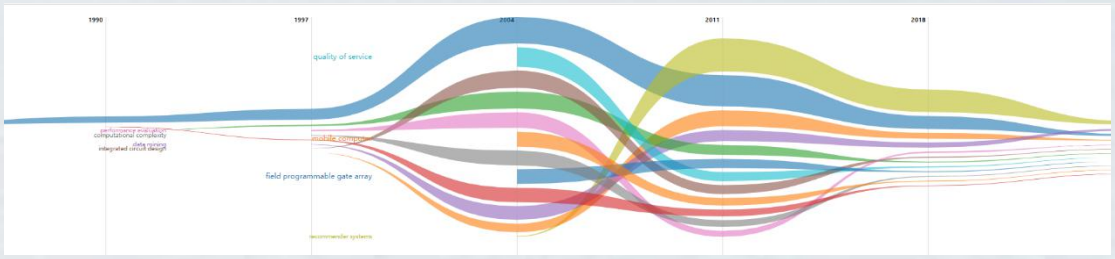


图 19 信息检索与推荐的热点趋势图

接着我们将挖掘到的各个时间段的前 10 个关键词列出，并按照出现次数降序呈现，如下表所示：

表 3 信息检索与推荐发展各时期的关键词表

1970-1979	1980-1989	1990-1999	2000-2009	2010-2019
Data Structure	Document Retrieval	Graph Theory	Data Mining	Social Network
Retrieval System	Information System	Structure Analysis	Computer Complex	Collaborate Filter
Document Retrieval	User Interface	Information System	Performance Evaluation	Search Engine
Data Manage	Software Engine	Query Language	Mobile Computer	Machine Learning
Information Storage	Artificial Intelligence	Document Retrieval	Embed System	Data Mining
Query Language	Document Cluster	Information Technology	Search Engine	Social Media
Social Science	Knowledge Base	Database Manage System	Formal Verification	Matrix Factor
Question Answer System	Retrieval System	Neural Net	Data visual	Text Mine
Information Science	Expert System	User Interface	User Interface	Wed Search
Data Base	Nature Language	Data Structure	Software	Language Model

	Process		Architecture	

从关键词表可以看出，User Interface（用户界面）和 Document Retrieval（文档检索）一直是信息检索与推荐领域的发展关键词，出现了三次。其中，User Interface（用户界面）在 1980-1989 年间出现的次数为 40；在 1990-1999 年间出现的次数为 38；在 2000-2009 年间出现的次数为 380，即用户界面的研究总体上呈递增趋势。而 Document Retrieval（文档检索）在 1970-1979 年间出现的次数为 6；在 1980-1989 年间出现的次数为 74；在 1990-1999 年间出现的次数为 63，即关于文档检索的研究总体上呈递减趋势。多媒体内容的出现对纯文档检索方法提出了新的挑战。

出现两次的关键词有 6 个，分别为 Data Structure（数据结构）、Query Language（查询语言）、Retrieval System（检索系统）、Information System（信息系统）、Data Mining（数据挖掘）、Search Engine（搜索引擎）。

其中，Data Structure（数据结构）和 Query Language（查询语言）首次出现都是在 1970-1979 年间；第二次出现在 1990-1999 年间，即它们的研究热度成衰减趋势。

Retrieval System（检索系统）和 Information System（信息系统）是兴起时间较早，前者出现在 1970-1979 和 1980-1989 年间；后者出现在 1980-1989 和 1990-1999 年间；而 Data Mining（数据挖掘）和 Search Engine（搜索引擎）是近年来兴起的，出现在 2000-2009 和 2010-2019 年间，并且关于它们的研究热度不减。

此外还值得注意，在 2010-2019 年间 Social Network（社交网络）和 Social Media（社交媒体）这两个关键词的出现。21 世纪以来，随着网络技术的发展，信息检索与推荐的研究也逐渐呈现社交属性。

通过对 1970-2019 年间信息检索与推荐领域有关论文的挖掘，总结出二十多年来，信息检索与推荐的领域关键词主要集中在 User Interface（用户界面）、Document Retrieval（文档检索）、Search Engine（搜索引擎）、Query Language（查询语言）等领域。

5.2 技术预见

研究者根据信息检索与推荐领域近十年的相关论文，利用大数据分析、机器学习、人工智能等技术手段，建立算法模型及研发 demo 系统，分析挖掘出该领域的技术发展热点。

技术预见图中点的大小表示该技术的热点（主要由相关论文数量的多少决定，相关论文越多，热度越高，点越大），各技术之间的连线表示 2 个技术关键词同时在 N 篇论文中出现过（当前 N 的取值为 5）。

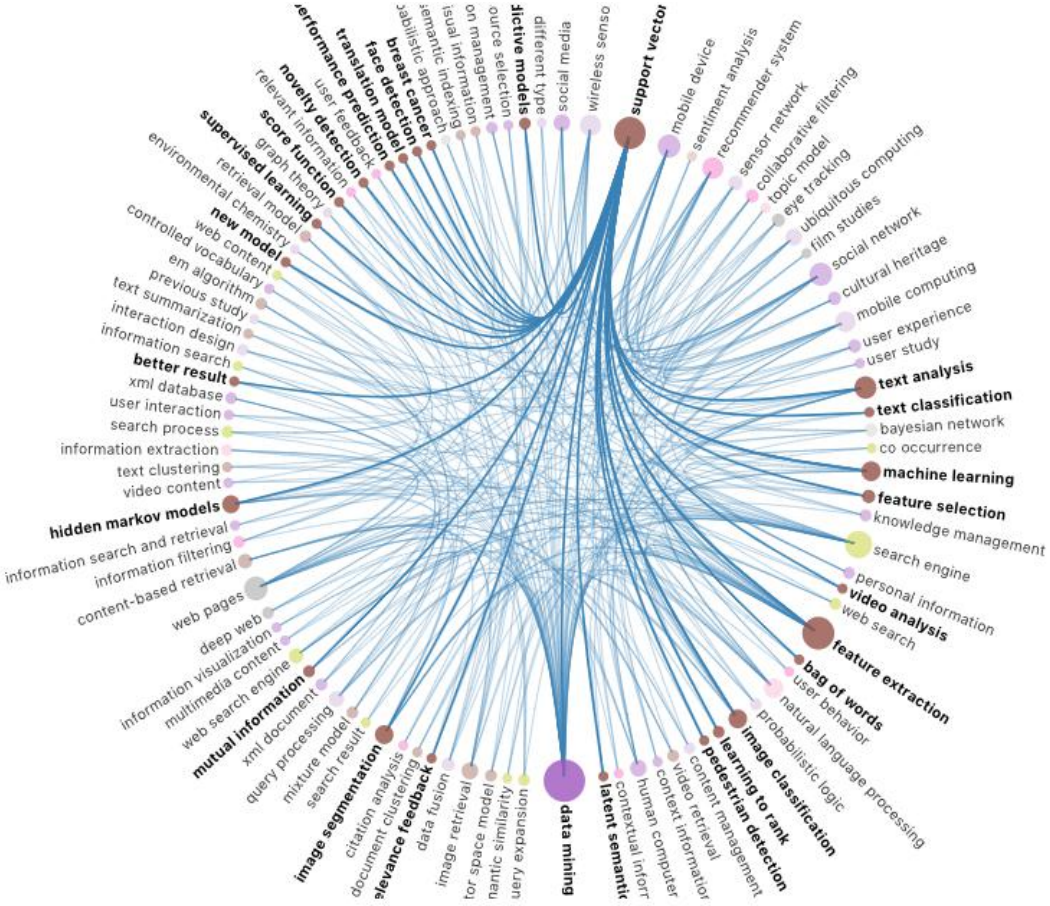


图 20 信息检索技术预见图

根据信息检索技术预见图（图 20），可以得出信息检索领域相关度最高的技术有 26 项，分别为：data mining（数据挖掘）、latent semantic（潜在语义）、learning to rank（学习排序）、image classification（图像分类）、bag-of-words model（词袋模型）、feature extraction（特征提取）、feature selection（特征选择）、machine learning（机器学习）和 text analysis（文本解析）等。

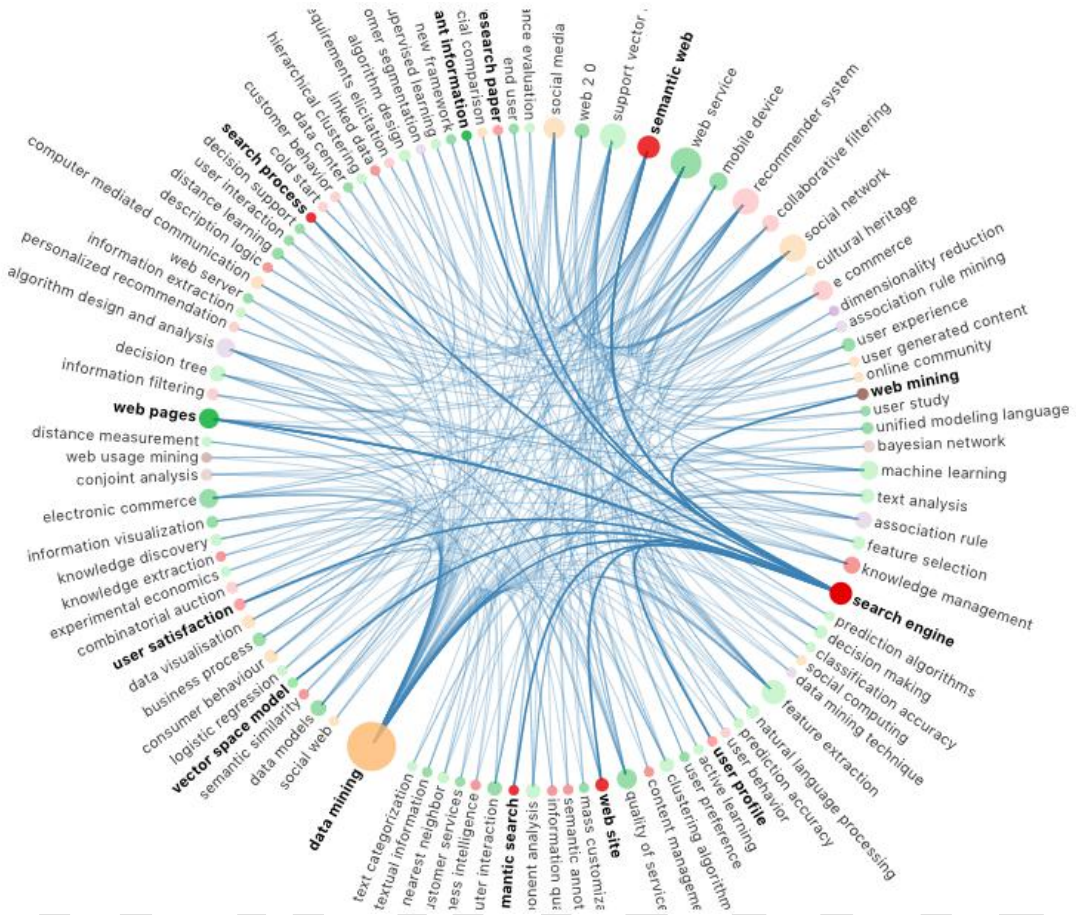


图 21 推荐系统技术预见图

根据信息推荐技术预见图（图 21），可以得出信息推荐领域相关度最高的技术有 13 项，分别为：search engine（搜索引擎）、web mining（web 挖掘）、semantic web（语义网络）、search process（检索过程）、web pages（网页页面）、user satisfaction（用户满意度）、vector space model（向量空间模型）、data mining（数据挖掘）、web site（网站站点）和 user profile（用户画像）等。

按照技术前沿度，可以列出相关的主要技术关键词，以及该技术历年的变化趋势（论文发表数量变化趋势），及重要代表性成果。具体如下图所示：

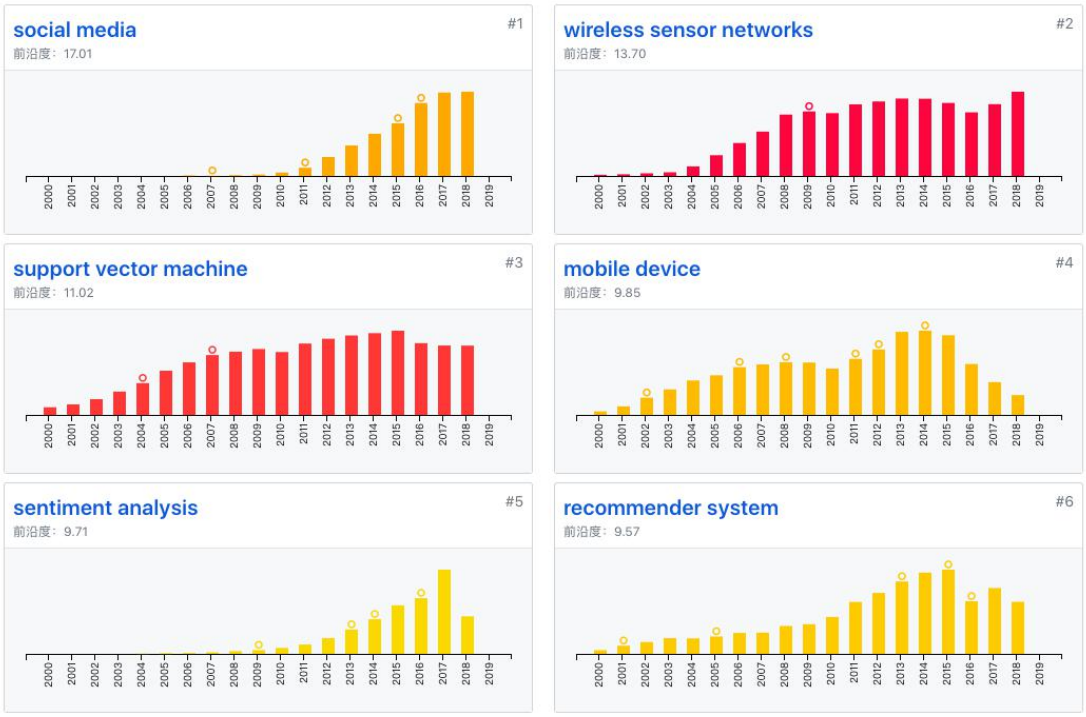


图 22 信息检索领域的六个技术关键词

从信息检索领域技术关键词图（图 22）可以看出，未来该领域的六个技术热点分别为：social media（社交媒体）、wireless sensor networks（无线传感器网络）、support vector machine（支持向量机）、mobile device（移动设备）、sentiment analysis（情感分析）和 recommender system（推荐系统）。

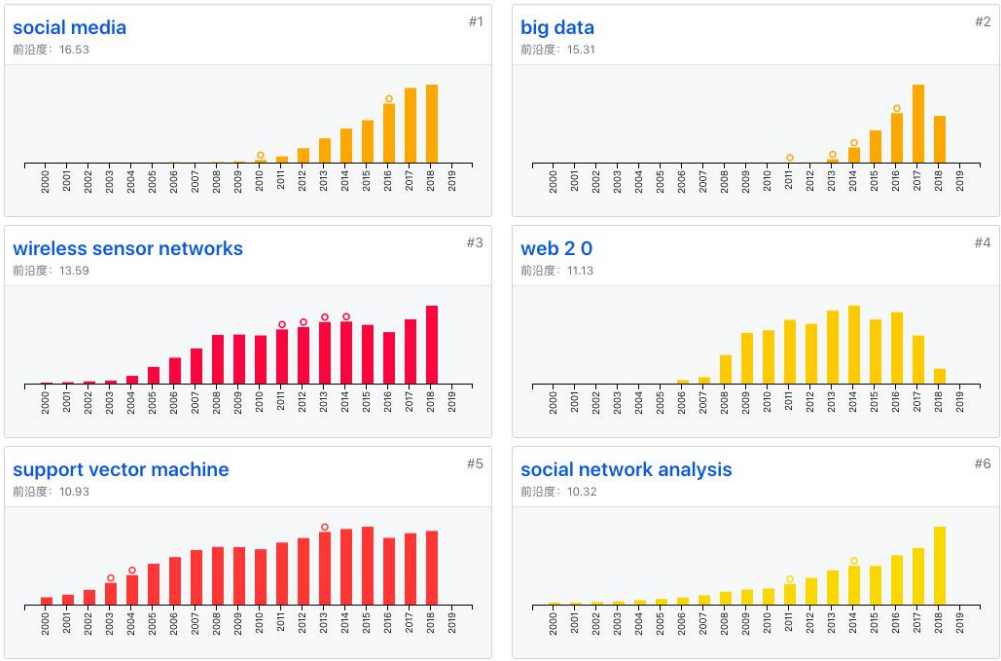


图 23 信息推荐领域的六个技术关键词

从信息推荐领域技术关键词图（图 23）可以看出，未来该领域的六个技术热点分别为：social media（社交媒体）、big data（大数据）、wireless sensor networks（无线传感器网络）、web2.0 support vector machine（支持向量机）和 social network analysis（社交网络分析）。

从总体上看检索技术发展迅速，但目前仍存在一些问题。例如搜索引擎存在缺陷且智能化程度低的状况，用户在使用搜索引擎检索信息时，经常会搜索到一些毫不相干的内容，影响了搜索体验；有些网站质量差，栏目设置很是混乱，对信息资源的分类组织上都存在着混乱状况，类目划分标准不合理，甚至可能出现同时用两个或两个以上标准划分的现象；综合性搜索引擎提供大众化服务较多，而个性化服务很少，它们没有有效的手段理解用户准确的个性化信息需求，不能提供长期的主动的信息服务。

虽然推荐系统已经成功运用于很多大型系统及网站，但是在当前大数据的时代背景下，推荐系统不仅面临数据稀疏、冷启动、兴趣偏见等传统难题，还面临由大数据引起的更多、更复杂的实际问题。

为了解决这些问题，信息检索在未来必须具有能及时挖掘新信息和及时链接新增的信息、多途径检索等功能；未来网络信息资源在组织分类上需要制定一个统一的分类标准，规范网络术语，提高资源共享的程度；每个人对信息的需求也不再满足于标准化、单一化的大众需求，更加智能的融合 NLP 技术将更好地去理解用户显性的意图或者是隐性的意图；新的信息交互模式也是未来行业关注的焦点，交互式会话检索也会在业界越来越受关注，开发出更好的对其评测方法也是关注点之一。相信在众多信息专家努力下，在信息检索与推荐领域将取得更大的突破，人们可以获取更多丰富的信息资源。

参考文献

[1] Ricardo Baeza-yates, Berthier Ribeiro-Neto 著. 黄宣菁，张奇，邱锡鹏译. 《现代信息检索》. 机械工业出版社. 2012 年 10 月第 1 版.

[2] 金芳.浅谈信息检索与信息检索技术[J].晋图学刊,2001(03):22-24+49.

[3] 刘弈群、马少平、洪涛、刘子正. 《搜索引擎技术基础》. 清华大学出版社出版. 2010 年 7 月第 1 版.

[4] 详细分析推荐系统和搜索引擎的差异. https://blog.csdn.net/cserchen /article/details/50422553

[5] Manning C, Raghavan P, Schütze H. Introduction to information retrieval[J]. Natural Language Engineering, 2010, 16(1): 100-103.

[6] Salton G, Fox E A, Wu H. Extended Boolean information retrieval[R]. Cornell University, 1982.

[7] Ogawa Y, Morita T, Kobayashi K. A fuzzy document retrieval system using the keyword connection matrix and a learning method[J]. Fuzzy sets and systems, 1991, 39(2): 163-179.

[8] Cao, Z., Qin, T., Liu, T.-Y., Tsai, M.-F., & Li, H. (2007). Learning to rank: from pairwise approach to listwise approach. In Proceedings of the 24th international conference on Machine learning (pp. 129–136).

[9] 李金忠,刘关俊,闫春钢,蒋昌俊.排序学习研究进展与展望[J].自动化学报,2018,44(08):1345-1369.

[10] 孙建文. 基于深度学习的中文文档检索的应用[D].吉林大学,2015.

[11] Jones K S. Index term weighting[J]. Information storage and retrieval, 1973, 9(11): 619-633.

[12] Furnas G W, Deerwester S., Dumais S T, et al. Information retrieval using a singular value decomposition model of latent semantic structure[C]//Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1988: 465-480.

[13] Robertson S E, Jones K S. Relevance weighting of search terms[J]. Journal of the American Society for Information science, 1976, 27(3): 129-146.

[14] Rosenfeld R. Two decades of statistical language modeling: Where do we go from here?[J]. Proceedings of the IEEE, 2000, 88(8): 1270-1278.

[15] 查正军,郑晓菊.多媒体信息检索中的查询与反馈技术[J].计算机研究与发展,2017,54(06):1267-1280.

[16] 郭少友. 基于会话管理的 Web 即时信息检索研究[J]. 图书情报工作, 2009, 53(16).

[17] 臧劲松.人工智能在跨语言信息检索中的应用[J].计算机时代,2016(10):29-31+35.

[18] Ballesteros, L., & Croft, W. B. (1998). Resolving ambiguity for cross-language retrieval. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (pp. 64–71).

[19] 黄立威,江碧涛,吕守业,刘艳博,李德毅.基于深度学习的推荐系统研究综述[J].计算机学报,2018,41(07):1619-1647.

[20] 李伟. 基于关联规则 B2C 图书销售网站个性化推荐系统研究[D].对外经济贸易大学,2007.

[21] 陈春玮. 基于关联规则和神经网络分析的推荐系统的研究[D].杭州电子科技大学,2017.

[22] 王静. 基于关联规则的图书销售网站个性化推荐系统设计与实现[D].电子科技大学,2012.

[23] 姚婷. 基于协同过滤算法的个性化推荐研究[D].北京理工大学,2015.

[24] 孔维梁. 协同过滤推荐系统关键问题研究[D].华中师范大学,2013.

[25] 推荐系统-第四章-基于知识的推荐. http://blog.sina.com.cn/s/blog_7103b28a0102wlo1.html.

[26] Zhang Y, Chen X. Explainable Recommendation: A Survey and New Perspectives[J]. 2018.

[27] 协同过滤推荐算法的原理及实现. https://blog.csdn.net/yimingsilence/article/details/54934302.

[28] 网易云音乐推荐算法分析. https://zhuanlan.zhihu.com/p/63908049.

版权声明

主　　编：刘奕群　　唐　杰

责任编辑：刘　佳

编　　辑：唐丽杭　　景　晨

封面设计：边云风　　朴之贤