gitbook.cn

GitChat - 《自然语言处理研究报告》的文章

31 ～ 39 分

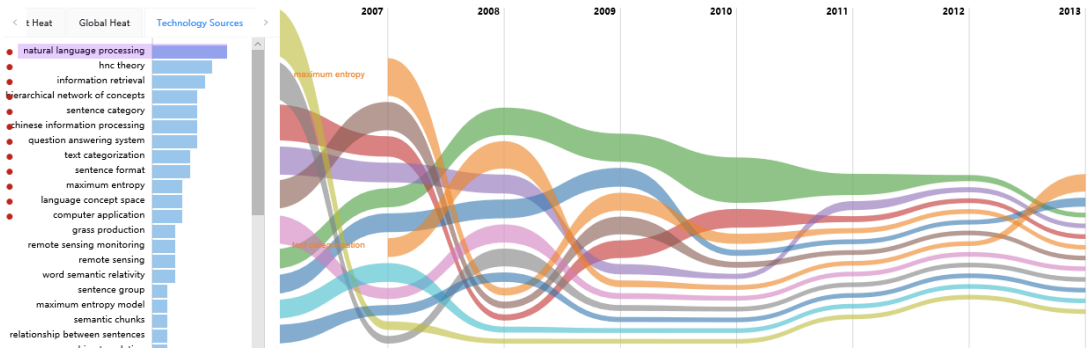


图 2 自然语言处理技术起源

自然语言处理的研究领域极为广泛，各种分类方式层出不穷，各有其合理性，我们按照中国中文信息学会2016年发布的《中文信息处理发展报告》，将自然语言处理的研究领域和技术进行以下分类，并选取其中部分进行介绍。

表1 自然语言处理技术分类及代表在学者

	技术	代表学者
基础技术	词法与句法分析	张民
	语义分析	周国栋、李军辉
	语篇分析	王厚峰、李素建

核心技术	语言认知模型	宗成庆
	语言知识表示与深度学习	黄萱菁、邱锡鹏
	知识图谱	李涓子
	文本分类与聚类	刘知远、涂存超
	自动文摘	万小军、姚金戈
	多模态信息处理	陈晓鸥
	信息抽取	孙乐、韩先培
	文字识别	刘成林
	语音技术	郑方、陶建华、王东
应用技术	信息推荐与过滤	王斌
	自动问答	赵军
	机器翻译	宗成庆、张家俊
	信息检索	马少平、刘奕群
	情感分析	黄民烈
	社会媒体处理	刘挺

2.1 自然语言处理基础技术

这一小节重点介绍自然语言处理的基础研究方面，自然语言的基础技术包括词汇、短语、句子和篇章级别的表示，以及分词、句法分析和语义分析等。重点选取词法、句法和语义分析来进行介绍。

词法分析的主要任务是词性标注和词义标注。词性是词汇的基本属性，词性标注就是在给定句子中判断每个词的语法范畴，确定其词性并进行标注。解决兼类词和确定未登录词的词性问题是标注的重点。进行词性标注通常有基于规则和基于统计的两种方法。一个多义词往往可以表达多个意义，但其意义在具体的语境中又是确定的，词义标注的重点就是解决如何确定多义词在具体语境中的义项问题。标注过程中，通常是先确定语境，再明确词义，方法和词性标注类似，有基于规则和基于统计的做法。

判断句子的句法结构和组成句子的各成分，明确它们之间的相互关系是句法分析的主要任务。句法分析通常有完全句法分析和浅层句法分析两种，完全句法分析是通过一系列的句法分析过程最终得到一个句子的完整的句法树。句法分析方法也分为基于规则和基于统计的方法，基于统计的方法是目前的主流方法，概率上下文无关文法用的较多。完全句法分析存在两个难点，一是词性歧义；二是搜索空间太大，通常是句子中词的个数n的指数级。浅层句法分析又叫部分句法分析或语块分析，它只要求识别出句子中某些结构相对简单的成分如动词短语、非递归的名词短语等，这些结构被称为语块。一般来说，浅层语法分析会完成语块的识别和分析，语块之间依存关系的分析两个任务，其中语块的识别

和分析是浅层语法分析的主要任务。

语义分析是指根据句子的句法结构和句子中每个实词的词义推导出来能够反映这个句子意义的某种形式化表示，即将人类能够理解的自然语言转化为计算机能够理解的形式语言。句子的分析与处理过程，有的采用“先句法后语义”的方法，但“句法语义一体化”的策略还是占据主流位置。语义分析技术目前还不是十分成熟，运用统计方法获取语义信息的的研究颇受关注，常见的有词义消歧和浅层语义分析。

自然语言处理的基础研究还包括语用语境和篇章分析。语用是指人对语言的具体运用，研究和分析语言使用者的真正用意，它与语境、语言使用者的知识涵养、言语行为、想法和意图是分不开的，是对自然语言的深层理解。情景语境和文化语境是语境分析主要涉及的方面，篇章分析则是将研究扩展到句子的界限之外，对段落和整篇文章进行理解和分析。

除此之外，自然语言的基础研究还涉及词义消歧、指代消解、命名实体识别等方面的研究。

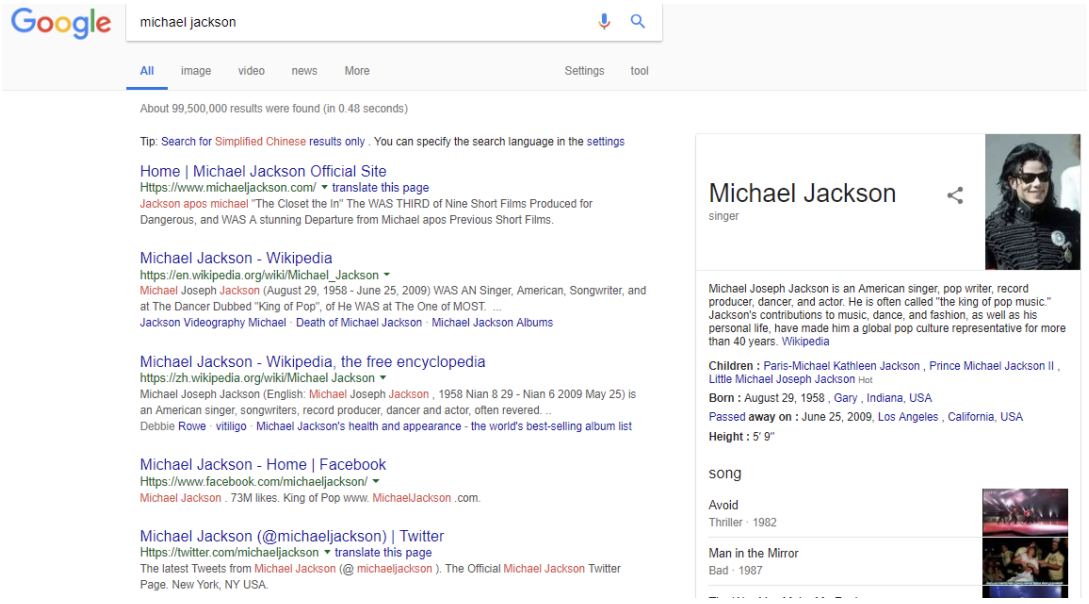
2.2 自然语言处理核心技术

自然语言处理核心技术包括知识图谱、文本分类与聚类、自动文摘、自动问答、信息抽取、文字识

别、语音技术、信息推荐与过滤、多模态信息处理等。

2.2.1 知识图谱

2012年5月，Google 推出 Google 知识图谱，并将其应用在搜索引擎中增强搜索能力，改善用户搜索质量和搜索体验，这是“知识图谱”名称的由来，也标志着大规模知识图谱在互联网语义搜索中的成功应用。搜索关键词，google 会在右侧给出与关键词相关的搜索结果。



知识图谱，是为了表示知识，描述客观世界的概念、实体、事件等之间关系的一种表示形式。这一概念的起源可以追溯至语义网络——提出于20世纪五六十年代的一种知识表示形式。语义网络由许多个节点和边组成，这些节点和边相互连接，节点表示的是概念或对象，边表示各个节点之间的关系，

如下图。

图 3 语义网络示意图

知识图谱在表现形式上与语义网络比较类似，不同的是，语义网络侧重于表示概念与概念之间的关系，而知识图谱更侧重于表述实体之间的关系。现在的知识网络被用来泛指大规模的知识库，知识图谱中包含的节点有以下几种：

实体：指独立存在且具有某种区别性的事物。如一个人、一种动物、一个国家、一种植物等。具体的事物就是实体所代表的内容，实体是知识图谱中的最基本元素，不同的实体间有不同的关系。

语义类：具有同种特性的实体构成的集合，如人类、动物、国家、植物等。概念主要指集合、类别、对象类型、事物的种类，例如人物、地理等。

内容：通常是实体和语义类的名字、描述、解释等，变现形式一般有文本、图像、音视频等。

属性（值）：主要指对象指定属性的值，不同的属性类型对应于不同类型属性的边。

关系：在知识图谱上，表现形式是一个将节点（实

体、语义类、属性值）映射到布尔值的函数。

除语义网络之外，70年代的专家系统以及 Tim Berners Lee 提出的语义网和关联数据都可以说是知识图谱的前身。

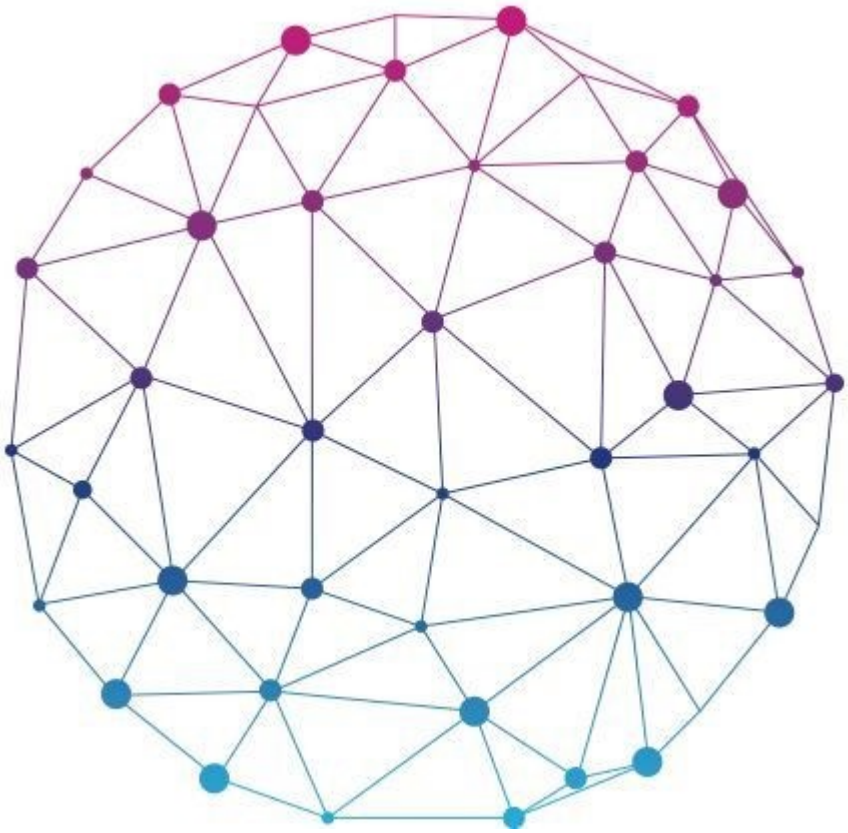


图 4 知识图谱示意图

知识图谱表示、构建和应用涉及很多学科，是一项综合的复杂技术。知识图谱技术既涉及自然语言处理中的各项技术，从浅层的文本向量表示、到句法和语义结构表示被适用于资源内容的表示中，分词和词性标注、命名实体识别、句法语义结构分析、指代分析等技术被应用于自然语言处理中。同时，知识图谱的研究也促进了自然语言处理技术的研究，基于知识图谱的词义排岐和语义依存关系分析

等知识驱动的自然语言处理技术得以建立。

2.2.2 自动问答

自动问答是指利用计算机自动回答用户所提出的问题以满足用户知识需求的任务。问答系统是信息服务的一种高级形式，系统反馈给用户的不再是基于关键词匹配排序的文档列表，而是精准的自然语言答案，这和搜索引擎提供给用户模糊的反馈是不同的。在自然语言理解领域，自动问答和机器翻译、复述和文本摘要一起被认为是验证机器是否具备自然理解能力的四个任务。

自动问答系统在回答用户问题时，首先要正确理解用户所提出的问题，抽取其中关键的信息，在已有的语料库或者知识库中进行检索、匹配，将获取的答案反馈给用户。这一过程涉及了包括词法句法语义分析的基础技术，以及信息检索、知识工程、文本生成等多项技术。传统的自动问答基本集中在某些限定专业领域，但是伴随着互联网的发展和大规模知识库语料库的建立，面向开放领域和开放性类型问题的自动问答越来越受到关注。根据目标数据源的不同，问答技术大致可以分为检索式问答、社区问答以及知识库问答三种。检索式问答和搜索引擎的发展紧密联系，通过检索和匹配回答问题，推理能力较弱。社区问答是 web2.0 的产物，用户生

成内容是其基础，Yahoo！Answer、百度知道等是典型代表，这些社区问答数据覆盖了大量的用户知识和用户需求。检索式问答和社区问答的核心是浅层语义分析和关键词匹配，而知识库问答则正在逐步实现知识的深层逻辑推理。

纵观自动问答发展历程，基于深度学习的端到端的自动问答将是未来的重点关注，同时，多领域、多语言的自动问答，面向问答的深度推理，篇章阅读理解以及对话也会在未来得到更广阔的发展。

2.2.3 自动文摘

自动文摘是运用计算机技术，依据用户需求从源文本中提取最重要的信息内容，进行精简、提炼和总结，最后生成一个精简版本的过程。生成的文摘具有压缩性、内容完整性和可读性。

从1955年 IBM 公司 Luhn 首次进行自动文摘的实验至今的几十年中，自动文摘经历了基于统计的机械式文摘和基于意义的理解式文摘两种。机械式方法简单容易实现，是目前主要被采用的方法，但是结果不尽如人意。理解式文摘是建立在对自然语言的理解的基础之上的，接近于人提取摘要的方法，难度较大。但是随着自然语言处理技术的发展，理解式文摘有着长远的前景，应用于自动文摘的方法也会越来越多。

自动文摘的分类方法多种多样，下表进行简单梳理：

表2 自动文摘分类

分类依据	类别		
摘要功能	指示摘要	信息摘要	评价摘要
与原文档关系	抽取（extraction）		摘要（abstraction）
对象	单文档摘要		多文档摘要
基于用户类型	主题摘要		普通摘要
机器学习角度	有指导的摘要		无指导的摘要

作为解决当前信息过载的一项辅助手段，自动文摘技术的应用已经不仅仅限于自动文摘系统软件，在信息检索、信息管理等各领域都得到了广泛应用。同时随着深度学习等技术的发展，自动文摘也出现了许多新的研究和领域，例如多文本摘要、多语言摘要、多媒体摘要等。

2.2.4 社会计算

社会计算也称计算社会学，是指在互联网的环境下，以现代信息技术为手段，以社会科学理论为指

导，帮助人们分析社会关系，挖掘社会知识，协助社会沟通，研究社会规律，破解社会难题的学科。社会计算是社会行为与计算系统交互融合，是计算机科学、社会科学、管理科学等多学科交叉所形成的研究领域。它用社会的方法计算社会，既是基于社会的计算，也是面向社会的计算。

社会媒体是社会计算的主要工具和手段，它是一种在线交互媒体，有着广泛的用户参与性，允许用户在线交流、协作、发布、分享、传递信息、组成虚拟的网络社区等等。近年来，社会媒体呈现多样化的发展趋势，从早期的论坛、博客、维基到风头正劲的社交网站、微博和微信等，正在成为网络技术发展的热点和趋势。社会媒体文本属性特点是其具有草根性，字数少、噪声大、书写随意、实时性强；社会属性特点是其具有社交性，在线、交互。它赋予了每个用户创造并传播内容的能力，实施个性化发布，社会化传播，将用户群体组织成社会化网络，目前典型的社会媒体是Twitter和Facebook，在我国则是微博和微信。社会媒体是一种允许用户广泛参与的新型在线媒体，通过社会媒体用户之间可以在线交流，形成虚拟的网络社区，构成了社会网络。社会网络是一种关系网络，通过个人与群体及其相互之间的关系和交互，发现它们的组织特点、行为方式等特征，进而研究人群的社会结构，以利于他们之间的进一步共享、交流与协作。

社会计算应用广泛，近年来围绕社会安全、经济、工程和军事领域得到了长足发展。金融市场采用社会计算方法探索金融风险和危机的动态规律，例如美国圣塔菲研究所建立了首个人工股票市场的社会计算模型。许多发达国家都在政府资助下开展了研究项目，例如美国的ASPEN，欧盟的 EURACE 等，并且在国家相应的经济政策制定中发挥着越来越重要的作用。通过社交媒体来把握舆情、引导舆论也是社会计算在社会安全方面发挥的一个重要作用。军事方面，许多国家更是加大投入力度扶持军事信息化的发展。

2.2.5 信息抽取

信息抽取技术可以追溯到20世纪60年代，以美国纽约大学开展的 Linguish String 项目和耶鲁大学 Roger Schank 及其同时开展的有关故事理解的研究为代表。信息抽取主要是指从文本中抽取出特定的事实信息，例如从经济新闻中抽取新发布产品情况如公司新产品名、发布时间、发布地点、产品情况等，这些被抽取出来的信息通常以结构化的形式直接存入数据库，可以供用户查询及进一步分析使用，为之后构建知识库、智能问答等提供数据支撑。

信息抽取和上文提到的信息检索关系密切，但是二

者之间仍存在着很大的不同。首先是二者要实现的功能不同，信息检索是要从大量的文档中找到用户所需要的文档，信息抽取则是用在文本中获取用户感兴趣或所需要的事实信息。其次是二者背后的处理技术也不同，信息检索依靠的主要是以关键字词匹配以及统计等技术，不需要对文本进行理解和分析，而信息则需要利用自然语言处理的技术，包括命名实体识别、句法分析、篇章分析与推理以及知识库等，对文本进行深入理解和分析后才能完成信息抽取工作。除了以上的不同之外，信息检索和信息抽取又可以相互补充，信息检索的结果可以作为信息抽取的范围，提高效率，信息抽取用于信息检索可以提高检索质量，更好地满足用户的需求。

信息抽取技术对于构建大规模的知识库有着重要的意义，但是目前由于自然语言本身的复杂性、歧义性等特征，而且信息抽取目标知识规模巨大、复杂多样等问题，使得信息抽取技术还不是很完善。但我们相信，在信息抽取技术经历了基于规则的方法、基于统计的方法、以及基于文本挖掘的方法等一系列技术演变之后，随着 web、知识图谱、深度学习的发展，可以为信息抽取提供海量数据源、大规模知识资源，更好地机器学习技术，信息抽取技术的问题会得到进一步解决并有长足的发展。

2.3 自然语言处理应用技术

自然语言处理应用技术包括机器翻译、信息检索、情感分析、社会媒体处理等。

2.3.1 机器翻译

机器翻译（Machine Translation）是指运用机器，通过特定的计算机程序将一种书写形式或声音形式的自然语言，翻译成另一种书写形式或声音形式的自然语言。机器翻译是一门交叉学科（边缘学科），组成它的三门子学科分别是计算机语言学、人工智能和数理逻辑，各自建立在语言学、计算机科学和数学的基础之上。

机器翻译的方法总体上可以分为基于理性的研究方法和基于经验的研究方法两种。

所谓“理性主义”的翻译方法，是指由人类专家通过编撰规则的方式，将不同自然语言之间的转换规律生成算法，计算机通过这种规则进行翻译。这种方法理论上能够把握语言间深层次的转换规律，然而理性主义方法对专家的要求极高，不仅要求其了解源语言和目标语言，还要具备一定的语言学知识和翻译知识，更要熟练掌握计算机的相关操作技能。这些因素都使得研制系统的成本高、周期长，面向小语种的翻译更是人才匮乏非常困难。因此，翻译知识和语言学知识的获取成为基于理性的机器翻译方法所面临的主要问题。

所谓“经验主义”的翻译方法，指的是以数据驱动为基础，主张计算机自动从大规模数据中学习自然语言之间的转换规律。由于互联网文本数据不断增长，计算机运算能力也不断加强，以数据驱动为基础的统计翻译方法逐渐成为机器翻译的主流技术。但是同时统计机器翻译也面临诸如数据稀疏、难以设计特征等问题，而深度学习能够较好的缓解统计机器翻译所面临的挑战，基于深度学习的机器翻译现在正获得迅速发展，成为当前机器翻译领域的热点。

机器翻译技术较早的被广泛应用在计算机辅助翻译软件上，更好地辅助专业翻译人员提升翻译效率，近几年机器翻译研究发展更为迅速，尤其是随着大数据和云计算技术的快速发展，机器翻译已经走进人们的日常生活，在很多特定领域为满足各种社会需求发挥了重要作用。按照媒介可以将机器翻译分为文本翻译、语音翻译、图像翻译以及视频和 VR 翻译等。

目前，文本翻译最为主流的工作方式依然是以传统的统计机器翻译和神经网络翻译为主。Google、Microsoft 与国内的百度、有道等公司都为用户提供了免费的在线多语言翻译系统。将源语言文字输入其软件中，便可迅速翻译出目标语言文字。Google 主要关注以英语为中心的多语言翻译，百度则关注以英语和汉语为中心的多语言翻译。另外，即时通

讯工具如 Googletalk、Facebook 等也都提供了即时翻译服务。速度快、成本低是文本翻译的主要特点，而且应用广泛，不同行业都可以采用相应的专业翻译。但是，这一翻译过程是机械的和僵硬的，在翻译过程中会出现很多语义语境上的问题，仍然需要人工翻译来进行补充。

语音翻译可能是目前机器翻译中比较富有创新意思的领域，吸引了众多资金和公众的注意力。亚马逊的 A lexa、苹果的 Siri、微软的 Cortana 等，我们越来越多的通过语音与计算机进行交互。应用比较好的如语音同传技术。同声传译广泛应用于国际会议等多语言交流的场景，但是人工同传受限于记忆、听说速度、费用偏高等因素门槛较高，搜狗推出的机器同传技术主要在会议场景出现，演讲者的语音实时转换成文本，并且进行同步翻译，低延迟显示翻译结果，希望能够取代人工同传，实现不同语言人们低成本的有效交流。科大讯飞、百度等公司在语音翻译方面也有很多探索。如科大讯飞推出的“讯飞语音翻译”系列产品，以及与新疆大学联合研发的世界上首款维汉机器翻译软件，可以准确识别维吾尔语和汉语，实现双语即时互译等功能。

图像翻译也有不小的进展。谷歌、微软、Facebook 和百度均拥有能够让用户搜索或者自动整理没有识别标签的照片的技术。图像翻译技术的进步远不局限于社交类应用。医疗创业公司可以利用计算机阅

览 X 光照片、MRI（核磁共振成像）和 CT（电脑断层扫描）照片，阅览的速度和准确度都将超过放射科医师。而且更图像翻译技术对于机器人、无人机以及无人驾驶汽车的改进至关重要，福特、特斯拉、Uber、百度和谷歌均已在上路测试无人驾驶汽车的原型。

除此之外还有视频翻译和 VR 翻译也在逐渐应用中，但是目前的应用还不太成熟。

2.3.2 信息检索

信息检索是从相关文档集合中查找用户所需信息的过程。先将信息按一定的方式组织和存储起来，然后根据用户的需求从已经存储的文档集合当中找出相关的信息，这是广义的信息检索。信息检索最早提出于20世纪50年代，90年代互联网出现以后，其导航工具——搜索引擎可以看成是一种特殊的信息检索系统，二者的区别主要在于语料库集合和用户群体的不同，搜索引擎面临的语料库是规模浩大、内容繁杂、动态变化的互联网，用户群体不再是具有一定知识水平的科技工作者，而是兴趣爱好、知识背景、年龄结构差异很大的网民群体。

信息检索包括“存”与“取”两个方面，对信息进行收集、标引、描述、组织，进行有序的存放是“存”。按照某种查询机制从有序存放的信息集合（数据

库）中找出用户所需信息或获取其线索的过程是“取”。信息检索的基本原理是将用户输入的检索关键词与数据库中的标引词进行对比，当二者匹配成功时，检索成功。检索标识是为沟通文献标引和检索关键词而编制的人工语言，通过检索标识可以实现“存”“取”的联系一致。检索结果按照与提问词的关联度输出，供用户选择，用户则采用“关键词查询+选择性浏览”的交互方式获取信息。

以谷歌为代表的“关键词查询+选择性浏览”交互方式，用户用简单的关键词作为查询提交给搜索引擎，搜索引擎并非直接把检索目标页面反馈给用户，而是提供给用户一个可能的检索目标页面列表，用户浏览该列表并从中选择出能够满足其信息需求的页面加以浏览。这种交互方式对于用户来说查询输入是简单的事，但机器却难以通过简单的关键词准确的理解用户的真正查询意图，因此只能将有可能满足用户需求的结果集合以列表的形式提供给用户。

目前互联网是人们获取信息的主要来源，网络上存放着取之不尽、用之不竭的信息，网络信息有着海量、分布、无序、动态、多样、异构、冗余、质杂、需求各异等特点。人们不再满足于当前的搜索引擎带来的查询结果，下一代搜索引擎的发展方向是个性化（精确化）、智能化、商务化、移动化、社区化、垂直化、多媒体化、实时化等。

2.3.3 情感分析

情感分析又称意见挖掘，是指通过计算技术对文本的主客观性、观点、情绪、极性的挖掘和分析，对文本的情感倾向做出分类判断。情感分析是自然语言理解领域的重要分支，涉及统计学、语言学、心理学、人工智能等领域的理论与方法。情感分析在一些评论机制的 app 中应用较为广泛，比如某酒店网站，会有居住过的客人的评价，通过情感分析可以分析用户评论是积极还是消极的，根据一定的排序规则和显示比例，在评论区显示。这个场景同时也适用于亚马逊、阿里巴巴等电商网站的商品评价。

除此之外，在互联网舆情分析中情感分析起着举足轻重的作用，话语权的下降和网民的大量涌入，使得互联网的声音纷繁复杂，利用情感分析技术获取民众对于某一事件的观点和意见，准确把握舆论发展趋势，并加以合理引导显得极为重要。

同时，在一些选举预测、股票预测等领域情感分析也逐渐体现着越来越重要的作用。