·学术科普·

漫威电影中的信息检索

Information retrieval in Marvel Cinematic Universe

王元卓　中国科学院计算技术研究所
沈英汉　中国科学院计算技术研究所
陆　源　竞技世界（北京）网络技术有限公司

在电影《复仇者联盟2：奥创纪元》中，有一个精彩的情节是当绿巨人浩克被绯红女巫迷惑心智后在城市中大肆破坏时，钢铁侠与他的专属智能大脑贾维斯对话，通过“浩克”这个关键词快速搜索实时新闻报道和视频中的相关信息，得出绿巨人的位置；钢铁侠及时赶到并启动了反浩克战甲，阻止了绿巨人的破坏行为。在这个情节中，智能大脑贾维斯依据钢铁侠给出的“浩克”这一关键词，从海量的新闻报道、网络视频数据中检索出相匹配的信息，并返回给钢铁侠，这里应用的就是信息检索技术，如图1所示。

图1　《复仇者联盟2》电影中的片段

信息检索技术这一名词在人们的生活中无处不在。可以说，只要是应用了搜索引擎的应用，都会有信息检索的影子。大家一定都用过百度、搜狗、谷歌、必应这些搜索引擎吧？在搜索引擎中搜索“浩克”，可以从海量数据中检索出浩克的基本信息、最新电影状况、相关演员动态等；在电商平台中搜索“浩克”，可以检索出与浩克相关的书籍、玩具、游戏等；在社交平台中搜索“浩克”，可以检索出与“浩克”一词相关的用户昵称、网友互动信息、短视频；等等。

信息检索技术极大地方便了人们的生活，能让人们快速定位感兴趣的信息，大大节省了人们的时间与精力。信息检索的基本原理是什么呢？首先，用户需要明确自己需要检索的信息是什么，并将检索信息

输入搜索引擎。例如，刚才提到《复仇者联盟2：奥创纪元》中的情节，钢铁侠需要检索的信息是“浩克在哪里”，这条信息中对应的知识可能就是“绿色”+“大块头”。搜索引擎首先会从全网信息中初步筛选出与浩克相关的信息（即包含“绿色”+“大块头”实体的信息）；钢铁侠的需求是明确浩克的地理位置，搜索引擎需要进一步从包含浩克的场景信息中筛选出浩克最有可能出现的实时地理位置信息并返回给钢铁侠。严谨地说，信息检索技术的基本原理可被概括为：从用户需求出发，对信息集合与需求集合进行匹配和选择，根据一定的线索与准则找出相关的信息。

信息检索技术的两种主流技术手段分别是关键词检索与语义检索。我们来聊聊这两种检索技术的实现方式。信息检索讲解图如图2所示。

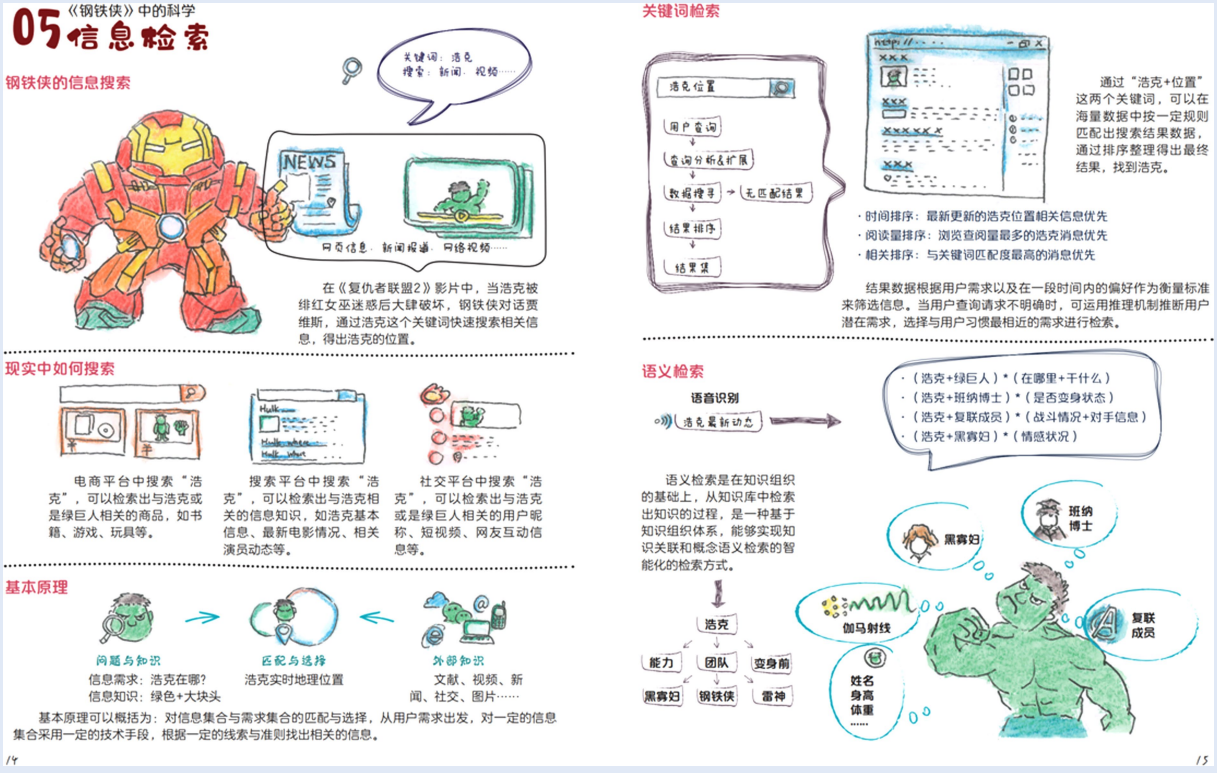


图 2　信息检索讲解图

（选自《科幻电影中的科学：科学家奶爸的AI手绘》）

关键词检索是指用户在搜索引擎的搜索框中键入自己要搜索信息的关键词，并进行检索的方式。在刚才的故事情节中，钢铁侠口述的检索信息中包含两个关键词，即“浩克”与“位置”。通过“浩克”与“位置”这两个关键词，搜索引擎从海量数据中按照关键词匹配规则筛选出搜索结果数据，并通过排序整理得出最有可能的结果。这种检索方式通过解析目标信息的字符，在语料库中搜索与内容相匹配的信息，具有查询信息范围大、正确率高、查全率高等优点，但是这种方式对包含海量数据的语料库的构建要求也较高。

一般来说，检索结果的排序方式有3种。第一种是按时间排序，最新更新的相关信息优先展示；第二种是按阅读量排序，浏览查阅次数最多的信息优先展示；第三种是相关排序，与关键词匹配度最高的信息优先展示。通常将用户需求以及用户在一段时间内的偏好作为衡量标准来对检索结果进行排序。当用户查询需求不明确时，可运用推理机制推断用户潜在需求，选择与用户习惯最相近的信息进行检索。

语义检索则是在概念体系的基础上，搜索引擎从知识库中检索出知识的过程。这是一种基于知识

图谱体系，能够实现知识间的关联，以及概念和概念语义检索的智能化检索方式。举例来说，在基于语义检索的搜索引擎中搜索“浩克”一词，搜索引擎不会通过文本的硬性匹配筛选数据，而是依据浩克这一实体对应的知识，检索与其相关的实体知识信息，如班纳博士、浩克的身高和体重，以及黑寡妇等复仇者联盟成员等。

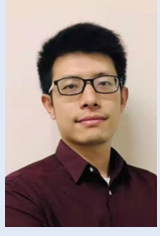
此外，语义检索可以将搜索的多个实体进行组合，并能够从实体组合中挖掘出更深层次的语义知识。如搜索“浩克”+“黑寡妇”，搜索结果为浩克的情感状态；搜索“浩克”+“钢铁侠”，搜索结果大部分为浩克和托尼·史塔克开发的反浩克装甲。相较于关键词检索中纯文本匹配的方式，语义检索更倾向于通过检索文本对应的知识，从知识库中检索出最有可能的结果。这种基于研究数据之间的关系的信息检索技术提高了数据检索能力，增强了自然语言的理解力，提升了查全率，但是也存在检索速度慢、查询复杂、耗费大量人工的缺点。

在本文中，我们了解了信息检索技术。它的两种主流技术手段分别是关键词检索与语义检索。更加智能化的信息检索系统模拟人类关于数据处理的思维过程和智能活动，实现知识检索、表示和推理，还可以为用户提供智能辅助决策。信息检索技术已被广泛应用到电商平台、新闻资讯、社交媒体、娱乐视频等软件中，可以帮助人们快速定位自己需要的信息，给人们的生活带来了极大便利。

作者简介

王元卓（1978– ），男，博士，中国科学院计算技术研究所研究员、博士生导师，中科大数据研究院院长，中国科普作家协会副理事长，中国计算机学会科学普及工作委员会主任，主要研究方向为大数据与人工智能。

沈英汉（1995– ），男，中国科学院计算技术研究所博士生，主要研究方向为社交知识图谱。



陆源（1990– ），男，现就职于竞技世界（北京）网络技术有限公司，从事数据产品工作，主要研究方向为大数据与社交网络。热心科普创作，科普畅销书《科幻电影中的科学：科学家奶爸的宇宙手绘》《科幻电影中的科学：科学家奶爸的AI手绘》作者之一。

中图分类号：TP399　　文献标志码：E　　doi: 10.11959/j.issn.2096-0271.2022079