人工智能在消费领域的研究现状分析——基于文本挖掘技术

晏瑶

西安外国语大学商学院，陕西西安，710128；

摘要：本文基于文本挖掘技术，采用文本分析的方法，基于 Anaconda3 可视化软件，通过八爪鱼数据搜集器在中国知网搜集 2012-2023 年以“人工智能”和“消费”为主题的中文文献。对数据拆分、分词后，进行数量、来源和类型统计，通过高频词提取、关键词分析、特征提取等方法追踪研究热点，概述“人工智能在消费”领域现状。分析显示，人工智能引发学者关注，领域热点包括智能化营销、智能算法，同时关注隐私安全、自动化研究，研究视角趋向多元化。

关键词：文本分析；人工智能；消费；研究热点

DOI：10.69979/3029-2700.24.9.012

1 引言

互联网进步带动人工智能发展，学者和企业纷纷将其引入消费场景。相关研究报告显示，在 2016 年时仅有 38%的企业引用了人工智能技术，而到了 2017 年就迅速增长到了 61%，人工智能技术的快速发展正深刻改变着消费领域的格局[1]。本文采用文本分析的方法，基于 Anaconda3 可视化软件，首先利用八爪鱼数据采集器，从中国知网数据库获取 2012 年-2023 年期间以“人工智能”、“消费”为主题的文献信息；其次对文献数据进行归纳整理，按照文章数量、发表期刊、发表时间等进行细分，分析整体分布情况；再次对文献数据（具体为摘要、关键词）进行文本分析，梳理当前的研究热点，并进行归纳总结，最后得出结论。

2 数据获取及预处理

2.1 数据来源

本文数据均来源于中国知网（CNKI）数据库，以“人工智能+消费”为搜索主题检索出所有文献，共获得 567 条检索记录，并通过使用 Excel 进行数据筛选从而删除与“人工智能”不相关的 34 篇文献，剩余可用 533 条检索信息，爬取的信息包括文献题名、题名链接、文献作者、摘要、关键词、发表时间、文献来源以及所属专题等详细信息[5][6]。

2.2 数据采集过程

本文的数据采集使用八爪鱼数据搜集器，在中国知网中搜集以人工智能为主题的中文文献，搜索关键词为“人工智能”、“消费”，得到 2012-2023 年以来的 500 多条文献数据，将每一条记录内的信息都抓取下来。主要抓取信息为：文献题名、题名链接、文献作者、作者链接、摘要、关键词、发表时间、文献来源以及所属专题。数据采集后，去除重复记录，存入 Excel 里，共得到 567 条记录。

2.3 数据预处理

由于在进行初步的文献数据爬取时使用了“人工智能”、“消费”为搜索关键词，因此某些含有“智能”、“消费”等内容，但与人工智能无关的文献信息也被检索了出来。如果在文献数据分析过程中继续使用这部分文献数据，会影响最终数据分析的有效性以及真实性，导致数据分析结果与实际情况存在较大偏差，所以在进行一系列的文本数据分析之前，应该提前对爬取到的数据进行预处理，过程如下：

2.3.1 过滤非相关数据

由于进行爬取时会不可避免获取到一些非相关数据，如下图中标注部分所示，需要将该部分内容删除。通过 Excel 里的“筛选”功能，设置不同的筛选条件，快速地剔除与“人工智能”、“消费”主题无关的内容，并保留有用信息，以达到较好的数据分析效果。原有 567 条文本信息，在剔除“书本编程”、“数字金融”等方面的不相关数据之后，剩余 533 条可用信息。

2.3.2 数据分列

由于数据采集最终得到的结果中，时间数据大多数包括了年、月、日三个维度，并用“—”分隔开，在 Excel 中存放在一列里。为获取文献发表的具体年份数据，需要将上述信息进行分解，采用 Excel 里的“分列”功能将其拆开，先在该列之后插入三个新列，选中“发表时

间”这一列，点击分列，选择“分隔符号”，最后得到文献发表时间的具体年份数据。

2.3.3 文本分词

文本分词，指将一句完整的话拆分成若干个词语的过程。在中文中，一句话内可能包含了动词、名词、以及部分没有实际意义的连词、助词等，为了便于对文本数据的处理与分析，采用文本分词技术将一段冗长、不易分析的文本转化为词语的组合，使语句文本变成结构化数据。文本分词对后续的分析奠定基础，本文在整体研究之前首先利用文本分析技术对爬取文献的摘要部分进行分词。

2.3.3 去停用词

在分析过程中，有一类词对分析数据没有实际意义，并且对情感极性没有影响，但是出现的频率却很高，对词频统计的结果产生较大的影响，例如“的”、“了”、“吧”等。去停用词是指将这些没有实际分析意义的词删除，减少文本中需要处理的词量，从而降低文本语言的复杂性，使文本初步处理结果更为简洁，增强文本分析的实际意义。一般来说，去停用词的方法是将词加入停用词表，将停用词表里的词语与分词后结果比对，删除相同词语。

因为每个文献摘要的信息里面通常都含有“本文”、“探讨”、“随着”、“研究”、“深入”等词语，这些对词频分析具有干扰作用，因此通过 stop words 进行删除。

2.3.4 近义词归并

另外，文本数据中还存在一些同义词，如“人工智能”、“AI”等，这些词在文中具有相同或类似的含义，但是由于呈现形式不同，往往会被分别统计，并不利于文本整体分析与内部关系探究，会造成最后分析效果呈现的繁杂与不准确，于是在近义词表里将这一类词进行归并。

3 研究现状分析

3.1 描述性统计分析

该部分内容针对现有文献的一些基础信息进行分析，本文以爬取到的 533 条文献记录为基础，展开以下描述性统计分析。

3.1.1 文章数量统计

对文章数量进行统计，选中“发表时间”列，通过插入数据透视表，值字段选择“计数”，通过对文献数量进行考察，发现在 2012-2023 年期间，相关文献数量增长速度变慢，其中 2020 年的文献数量达到最高，但是 2020 年之后，发表的文献数量开始减少。2020 年以前，文献数量始终保持增长，其中 2012-2016 年文献数量增长较少，直到 2017 年相关文献才超过 20 篇，而后 2017-2018 年文献数量增长最多。截至 2023 年 5 月，中国知网检索出的文献开始呈下降趋势，但是从整体的发展趋势来看，人工智能这一研究方向起步较早，从相对平缓的发展期，到近几年的快速增长趋势再到现在趋于平缓，这意味着该研究方向在当前阶段已经引起了学者们的高度关注，在未来仍待挖掘其他方面未知内容。

3.1.2 文章来源分布

选中文章“来源”列，拓展选区进行升序排序，然后通过分类汇总，得出前 15 个文章来源，其中前五位分别是西南财经大学、现代广告、中国市场、华南理工大学、营销界。

对“人工智能”相关文献的来源进行分析，发现文章出处最多的是西南财经大学，之后是数量相当的其他学校与期刊。总的来看，我们发现与人工智能这一话题相关的研究不仅与国家经济科技的发展有关，还与学者们所处的环境有关，例如不同学校对人工智能的感知不同，不同期刊对人工智能之一话题的感知也不同。从文章来源来看，许多学校与期刊来源的文献数量差距不是很大，也说明人工智能现在已经深入到社会环境各个方面，目前许多研究机构都对该领域研究有一定的基础积累。另一方面，目前发表的相关文献可以为后续研究提供丰富的资源和借鉴，也可以引领研究者探索新方向。

3.1.3 文献类型分布

对文章类型进行分析，通过分类汇总的方式得到各类文章数量，前五位分别为期刊、硕士、特色期刊、博士、辑刊。

从现有的文献分类来看，目前“人工智能”与“消费”领域的文献主要来自各个不同类型的期刊，另外则是硕士论文，二者的文献数量占绝大部分。另外，还有部分为博士论文，各类期刊论文为该领域的研究提供了资源借鉴，而博士论文的深入度为该领域的研究梳理了不同的发展历程与理论背景。

3.2 研究热点分析

3.2.1 高频词提取

为准确了解目前“人工智能+消费”领域的研究热点，了解学者们对该领域各特征研究的关注情况，因此对文章摘要中的特征词进行提取。通过对分词后的摘要进行词性标注可知，研究热点的特征词是由名词及动词

构成，因此对摘要文本中出现的名次及动词进行提取，由于所提取到的特征词过多，因此此处选取词频排名前 60 的词汇。提取结果如下表所示：

表 1 高频词

排名	特征词	词频	排名	特征词	词频
1	技术	857	31	信息	163
2	消费者	779	32	机器人	162
3	发展	745	33	责任	162
4	智能	582	34	模式	156
5	数据	494	35	用户	153
6	产品	411	36	产业	152
7	算法	408	37	自动	152
8	营销	385	38	体验	145
9	广告	359	39	提出	145
10	影响	293	40	零售	143
11	企业	284	41	歧视	139
12	金融	266	42	传播	138
13	科技	248	43	保护	137
14	汽车	248	44	基础	135
15	领域	246	45	数字	134
16	理论	243	46	价值	134
17	经济	239	47	提供	131
18	服务	234	48	场景	125
19	市场	216	49	需求	123
20	创新	213	50	平台	123
21	社会	210	51	侵权	123
22	互联网	202	52	感知	121
23	我国	200	53	系统	120
24	行业	197	54	品牌	119
25	传统	188	55	规制	118
26	带来	185	56	方式	117
27	时代	178	57	生活	114
28	驾驶	175	58	未来	114
29	风险	174	59	监管	111
30	法律	168	60	制度	110

由以上词频表不难看出，在“人工智能+消费”领域的相关研究中，大家都密切关注的是技术、消费者、发展、智能、数据等，而驾驶、法律、场景、体验等作为该领域衍生的话题也被部分学者所关注到，这说明“人工智能+消费”领域的相关研究现在已经开始转向多方位的研究，不单单只关注人工智能本身，也不单单只关注消费这一话题。

3.2.2 关键词分析

同时，对摘要文件进行关键词分[2][3]，得到以下关键词，如：消费者、智能、技术、算法、营销等，此处选取排名前 60 的词汇进行分析。从关键词分析可知，消费者、智能、技术、算法、营销等是目前研究中关注最多的，与高频词分析得到的结论相似。另外，从关键词来看，产品、机器人、服务、自动化也是目前研究的重点内容，说明当前“人工智能+消费”领域相关研究较为完善，研究维度也比较丰富。

3.2.3 特征提取

（1）文本数据拆分

将爬取到的文本内容信息进行数据清洗之后，将摘要内容进行分句处理，对分句后文本进行聚类[4]，采用 K-means 聚类分析，轮廓系数为：0.040465672368036394，将选取的摘要信息数据拆分成 4 个部分，每个部分围绕一个主题展开，聚类之后分别对每个类别的内容再进行共现分析。

此处共 4 个聚类：聚类 0 中，大多数语句都提到人工智能算法、智能算法、大数据算法等，该类别主要围绕“人工智能”的算法方面展开；聚类 1 中，语句涉及到的关键主题比其他聚类多，该类别主要围绕技术、智能、自动等方面内容展开；聚类 2 中，大多数语句都提到发展等，该类别主要围绕“人工智能”营销发展方面展开；聚类 3 中，大多数语句都提到消费者等，该类别主要围绕“人工智能+消费”领域中有关消费者方面的研究展开。

（3）特征词提取

对“人工智能+消费”领域的研究现状进行研究，为准确分析目前的研究热点与研究趋势，了解各项特征的被关注情况，对文本数据中的特征词进行提取。在前面几个步骤中，通过对文本进行分句处理后进行拆分得到 4 个聚类。由于所提取到的特征词较多，因此仅对每个聚类中词频排名显著多的词汇进行分析。第一个部分关键词为“算法”，第二个部分关键词为“技术”，第三个部分关键词为“发展”，第四个部分关键词为“消费者”，分别对 4 个部分进行分析以获得更准确的数据结果。

聚类 0 显示与人工智能研究密切相关的内容有“智能算法”，同时学者们在进行相关研究时还关注到有智能化引发的“歧视”问题，在消费领域，法律、规制、价格、反垄断等均被提及，在研究中受到一定关注。

聚类 1 显示在“人工智能”加“消费”的主题下，现有研究已经衍生到汽车领域，涵盖汽车的自动驾驶以及安全隐私问题，另外几个主要的研究重点是“智能化”、“广告”、“技术”、“营销”、“传播”，另外，现在关于该领域的研究还有部分专注于理论研究。

聚类 2 显示与“人工智能”相关研究密切相关的另

一特征为“发展”，技术、金融、大数据、产品等的发展普遍受到大众关注。

聚类 3 显示在消费领域，有关“人工智能”的研究大多数集中关注“消费者”，与消费者相关的服务、隐私保护、购买意愿、产品推广、广告设计等方面内容成为研究重点。

总的来看，目前在“人工智能+消费者”研究领域，研究的主题主要可以展开为四类，分别为“人工智能算法”，“人工智能技术”，“人工智能及其在营销领域发展历程”与“消费者研究”。人工智能算法方面涉及到法律与信息保护方面的内容，如大数据算法如何保护个人信息；人工智能带来的歧视方面的问题需要法律的规制，人工智能导致的在产品价格等方面的问题需要政府进行约束。人工智能技术方面涉及到自动驾驶、智能营销、创新科技、广告设计等方面的内容；如人工智能的发展带动汽车的无人驾驶技术的发展，同时产生责任分配的问题，智能营销引发智能客服、智能营销系统、智能语音等一系列问题，同时产生智能广告生成与广告安全的问题。消费者研究领域涉及到消费者意愿的相关研究，如对人工智能产品的接受意愿、被推荐产品时的接受度，以及影响消费者意愿与购买行为的因素研究[7]。

4 结论与展望

4.1 人工智能在消费领域的研究现状与热点概述

通过前文的描述性统计分析可以看出，2016 年之后，对人工智能与消费的相关研究快速增长，一方面对应了我国经济的发展，一方面也是我国互联网技术发展的体现。直到 2020 年，相关研究文献数量都在大幅增长，但是 2020 年以后，文献数量却开始减少，这其中不乏有 2020 年新冠疫情带来的影响，到 2023 年 5 月，中国知网检索出的文献数量仍比较少。从整体发展趋势来看，人工智能引入到消费领域进行的研究起步并不晚，这个意识已经较早出现，经历了平缓的发展期，再到高速发展，中间仅仅几年的时间。现在科技仍在不断发展，该研究方向在当前阶段已经引起了学者们的高度关注，同时，不难发现，不同环境下的研究者对人工智能及其在消费领域的应用感知是不同的，人工智能正在渗透到社会生活的方方面面。研究内容不仅关注到人工智能算法、人工智能技术、智能营销、消费者研究等方面，还衍生到许多更细化的方面，如隐私安全、社会歧视等，对企业来说，隐私与信息安全、消费者态度等是其在应用人工智能技术中应当重点关注的。

4.2 未来研究展望

信息技术的进步催生了人工智能的快速发展，使其成为我们生活中不可或缺的部分。在消费领域，人工智能的研究不仅关注技术本身，还扩展到自动化、信息安全等多个方面。同时，从研究重点以及研究数量的变化可以看出，社会发展具有整体性，当社会处于重大的外部不可抵挡事件中时，社会经济发展会受到一定影响，而它也会体现在某一个领域的研究中。尽管 2020 年以来，相关研究有所减少，但这可能是人工智能在消费领域研究和应用转型的起点，未来研究应更注重技术便利性与安全性。

参考文献

[1]刘龑龙,史冬梅,刘进长,唐莉,王金鹏.基于文献计量学的人工智能领域研究现状及热点分析[J].科技管理研究,2021,41(10):38-48.

[2]马秀麟,姜雪,贾玉娟.近十年面向人工智能教育研究的文献计量分析与探索[J].中国教育信息化,2022,28(08):35-46.

[3]严明.我国人工智能相关研究的期刊文献分析[J].西南民族大学学报(人文社科版),2019,40(11):229-234.

[4]徐畅,管开轩,宋昱晓,徐艳梅.文献计量视角下全球人工智能领域研究态势与热点分析[J].科技促进发展,2021,17(11):1968-1977.

[5]高娟,王静芬.基于 TDA 和 CiteSpace 的文献计量分析——以人工智能的应用研究为例[J].内蒙古科技与经济,2019(12):125-126+130.

[6]齐蒙.移动支付设计中用户消费行为研究进展——基于 Web of Science 2012—2020 年相关研究文献的分析[J].湖南包装,2020,35(06):17-21.

[7]王爱莲,冯睿.人工智能时代的市场营销研究综述[J].北方经贸,2021,No.443(10):55-57.

作者简介：晏瑶（2000—），女，汉族，贵州贵阳人，西安外国语大学商学院企业管理硕士研究生，主要研究方向为数字创新与品牌营销。