

数据到文本生成研究综述

曹娟¹, 龚隽鹏², 张鹏洲²

(1. 中国传媒大学 新媒体研究院, 北京 100024;

2. 中国传媒大学 理工学部, 北京 100024)

摘要: 机器新闻写作, 作为人工智能在传媒业的一种应用, 越来越受到学界和业界的关注, 目前主要用于体育、财经、气象地质和健康等领域。机器新闻写作的核心是自然语言生成技术, 而数据到文本生成是自然语言生成领域的典型技术, 是实现机器新闻写作的关键技术之一。为了更好地研究数据到文本生成技术并将其应用于机器新闻写作领域, 以内容选择和表层实现为重点, 梳理了近年来数据到文本生成的发展脉络, 并比较了基于规则和数据驱动两种研究方法, 归纳了不同领域的可用数据集, 总结了内在和外在两类评价方法, 分析了数据到文本生成技术当前存在的问题, 以及探讨了其未来可能的研究方向。

关键词: 数据到文本生成; 机器新闻写作; 自然语言生成; 内容选择; 表层实现; 神经网络

中图分类号: TP391

文献标识码: A

文章编号: 1673-629X(2019)01-0080-05

doi: 10.3969/j.issn.1673-629X.2019.01.017

Review of Data-to-text Generation

CAO Juan¹, GONG Jun-peng², ZHANG Peng-zhou²

(1. New Media Institute, Communication University of China, Beijing 100024, China;

2. Faculty of Science and Technology, Communication University of China, Beijing 100024, China)

Abstract: As an application of artificial intelligence in the media industry, robot journalism has attracted more and more attention from academia and industry. It is mainly used in sports, finance, meteorology, geology, health and other fields. The core of robot journalism is natural language generation, and data-to-text generation is a typical technology in the field of natural language generation, and also one of the key technologies to realize robot journalism. In order to study data-to-text generation technology better and apply it to the field of robot journalism, focusing on content selection and surface realization, we comb the development of data-to-text generation in recent years, compare two research methods based on rule and data-driven, sum up the available datasets in different areas, summarize the intrinsic and extrinsic evaluation methods, analyze the existing problems in data-to-text generation technology, and discuss the possible future research direction.

Key words: data-to-text generation; robot journalism; natural language generation; content selection; surface realization; neural network

0 引言

自从腾讯2015年推出Dream writer, 机器新闻写作开始受到国内研究者的关注, 并迅速成为学界和业界研究的热点。机器新闻写作是指基于数据分析和机器学习, 运用算法, 从可识别的数据中提取具有新闻价值的信息, 形成新闻报道角度, 自动选择语词样本、新闻报道模板生成新闻故事^[1]。国外的研究者称机器新闻为机器人新闻(robot journalism)或自动化新闻(automated journalism)。目前国外已经投入市场的产品

包括美国Automated Insights公司的Wordsmith和Narrative Science公司的Quill, 国内企业也做了一系列探索和尝试, 包括腾讯公司的Dream writer、新华社的快笔小新、今日头条的Xiaoming bot、第一财经的DT稿王、南方都市报的“小南”和广州日报的“阿同”^[2]。当前适合通过机器或算法进行的新闻写作, 一般是以各种数据、图表的引用和分析为基础的硬新闻, 具有明显的数据处理色彩, 主要用于财经、体育、气象地质和健康等领域^[3]。

收稿日期: 2018-02-12

修回日期: 2018-06-19

网络出版时间: 2018-11-15

基金项目: 国家科技支撑计划项目(2014BAK10B01)

作者简介: 曹娟(1990-), 女, 博士, CCF会员(44317G), 研究方向为机器新闻写作; 张鹏洲, 博士后, 研究员, CCF会员(07635S), 研究方向为内容管理、电视台整体化管理及其应用软件开发。

网络出版地址: <http://cnki.net/kcms/detail/61.1450.TP.20181114.1556.034.html>

机器新闻写作的核心在于自然语言生成(natural language generation, NLG)技术。自然语言生成中一个典型技术就是使用恰当而流畅的文本来描述结构化的数据,即数据到文本生成(data-to-text generation; data-to-document generation)。数据到文本生成可以归结为两大任务:说什么(what to say)和怎么说(how to say it)。说什么是从输入的数据中选择合适的子集用于表达,也叫做内容选择(content selection);怎么说就是用自然语言描述这个子集,也叫做表层实现(surface realization)^[4]。文中对近年来数据到文本生成的发展脉络和研究方法进行梳理,对已有数据集和评价方法进行总结,分析当前存在的问题并探讨其未来发展方向。

1 发展脉络和研究方法

数据到文本生成主要使用了基于规则(rule-based)的方法和数据驱动(data-driven)的方法。基于规则的方法,通常需要将内容选择和表层生成看作两个独立的子任务来完成;数据驱动的方法,可以单独用于内容选择,也可以将内容选择和表层生成看作一个整体来完成。下面介绍近年来数据到文本生成在这两种方法上的发展过程。

1.1 基于规则的方法

Sripada 等在 2001 年提出了针对时间序列数据的两阶段内容选择模型(two-stage model for content selection)^[5]。该模型基于人类专家对时间序列数据总结的观察,第一阶段构建数据集的定性概述,第二阶段结合实际数据生成总结。

Hallett 等在 2006 年针对医疗病史的总结提出一种内容选择方法^[6]。他们将一系列概念和事件联系在一起,在内容选择的过程中,将事件根据相关性聚在一起,假设小的集群不包含重要事件,因此在总结中只提到大的集群事件。依据总结的类型和长度,在基于规则的方式里决定内容的属性。

Turnertt 等在 2008 年使用决策树(decision tree)方法解决了在地理坐标参照(geo-referenced)数据描述领域中的内容选择问题^[7]。其中,树的叶子代表内容,节点代表事件,文本从叶子中的内容产生。在同样的领域,Thomas 等于 2012 年使用文档模式(document schemas)为盲人用户解决了地理坐标参照数据文本描述的文档规划(document plan)问题^[8]。其中,模式的选择受到空间数据分析的影响。

Gatt 等在 2009 年 BabyTalk 系统中使用了 Hallett 等在 2006 年提出的算法,用于内容规划(content plan),生成新生儿重症监护室数据的文本总结^[9]。这些数据包括传感器数据(心率、血压和血氧饱和度);

实验室结果和观察;事件如护士行为、医疗诊断和治疗等信息;自由文本。

Black 等在 2010 年为了帮助孩子解决复杂的沟通需求(complex communication needs)构建了一个工具,基于孩子可自编辑的传感数据使用 NLG 技术创造他们每天在学校的故事^[10]。输入的传感数据主要由孩子的位置、活动和与人或物体的交互构成。他们使用了无线射频识别技术(radio frequency identification, RFID),识别和监控位置和交互,用麦克风记录事件,还提供了可视化窗口。学校的老师和工作人员也可以访问孩子的活动信息。Tintarev 等于 2016 年进一步完善了该系统,根据位置、时间和语音对事件进行聚类分类,来决定叙述的内容,还使用规则定义意想不到或不平常的事件^[11]。

Banaee 等在 2013 年介绍了一种用于总结生理传感器数据例如心率和呼吸率的内容选择方法^[12]。从数据分析得到的抽象数据会被分成三种消息类型之一:全局信息、基于事件的消息和基于总结的消息,对于每个类型的消息都有一个单独的排名函数评估文本中消息的重要性,最后根据消息的重要性和事件之间的依赖性对消息进行排序。同年,Schneider 等介绍了跨学科 MIME 项目,即一个移动医疗监测系统,帮助进入医院前现场第一人和救护医生交接事务^[13]。他们使用 NLG 总结医疗传感器的数据和护理员的观察和操作随时生成文本交接报告。其中,内容选择模块结合语料分析和专家咨询获得的规则使用树列关联被选信息,类似于修辞结构理论^[14](rhetorical structure theory, RST)。

Soto 等在 2015 年描述了一种使用模糊集生成短天气预报的方法^[15]。方法中,内容选择部分由模糊算子进行操作,从所有可用数据中选出有用数据并转化成数据对象。最后,创建事件列表用于生成。

Gkatzia 等在 2016 年提出两种基于规则的方法实现天气预报的生成^[16]。第一种方法使用了 Kootval2008 年针对天气预报中不确定信息推荐的准则,第二种方法模拟了专家在天气预报中选择内容的方式。相比第一种方法映射到不确定性,第二种方法在语言解释上更加自然。

通过与专家合作或从专家生成的语料中获取知识是推导规则的主要方式,因此基于规则的方法通常适用于特定领域,生成的文本可读性较强,工业界大部分使用这种方式。但规则的数量也会随着领域复杂度的增加而增加,开发维护系统的开销可能会很大。

1.2 数据驱动的方法

数据驱动的方法也被称作可训练(trainable)的方法。尽管 NLG 使用数据驱动方法比 NLP 的其他子领

域起步晚,但数据驱动的方法已经在 NLG 中占据了主导地位。

2003 年 Duboue 和 McKeown 提出一种内容选择方法,从文本语料中自动学习内容选择规则和获取相关语义,并用于人物传记的短文本生成^[17]。他们把内容选择当作分类任务,目标是判定一个数据库条目是否应该出现在输出中。

2005 年 Barzilay 和 Lapata 提出一种协作内容选择方法 (collective content selection),从语料和相关数据库中自动学习内容选择规则,并用于足球赛事报道中^[18]。与 Duboue 和 McKeown 在 2003 年提出的方法不同的是,他们把内容选择看作协作分类问题,考虑了数据库条目之间的依赖性。

Liang 等在 2009 年解决了数据记录和给定文本描述子句匹配的问题,提出一种半隐马尔可夫 (hidden semi-Markov) 匹配生成模型,统一实现了分割文本到话语并关联话语到每个对应记录的任务^[19]。

Angeli 等在 2010 年提出一种将内容选择和表层生成统一且与领域无关的实现方法^[20]。该方法在 2009 年 Liang 等的基础上,加入了对数线性 (log-linear) 模型,将生成过程细化成一系列本地决策 (local decision),先选择事件记录,再选择记录属性,最后选择一系列属性对应的模板。

Konstas 等在 2012 年展示了将内容选择和表层生成统一的无监督且与领域不相关的模型^[21]。该模型没有将生成过程分割成本地决策,而是使用了概率上下文无关语法 (probabilistic context-free grammar, PCFG),全局地描述了输入数据的固有结构。该模型还用了超图 (hypergraph) 结构来获得最好的推导。

Kondadadi 等在 2013 年使用基于模板的统计 NLG 框架将内容选择和表层生成的任务联合成一个统计学习过程^[22]。其中,支持向量机 (support vector machine, SVM) 是构建该模型的主要方法。

Sowdaboina 等在 2014 年使用机器学习 (machine learning) 方法解决了对时序数据总结的内容选择问题^[23]。机器学习方法被用来学习产生文本总结的潜在规则,目的是更加接近人类生成文本总结的规则。

Gkatzia 等在 2014 年展示并对比了两种实现内容选择的可训练的方法^[24]。第一种使用多标签分类方法学习被选择的内容;第二种使用强化学习方法总结时序数据,内容选择被看作马尔可夫决策问题^[25]。

Mahapatra 等在 2016 年提出一种从表格形式的非文本数据实现统计自然语言生成的方法^[26]。该方法使用了多分区图 (multi-partite graphs) 用于天气预报的生成,每个分区由数据集中的每个属性创建,内容从图中有概率性地被选出。

近年来,深度学习在 NLG 中得到越来越多的关注。Mei 等在 2016 年提出一种端到端 (end-to-end) 的与领域无关的基于编解码 (encoder-decoder) 框架的神经网络模型^[27],其中用到了基于长短期记忆网络 (long short-term memory, LSTM) 的循环神经网络 (recurrent neural network, RNN)。Lebret 等在 2016 年介绍了一种建立在文本生成的条件神经语言模型 (conditional neural language models) 基础上的神经模型,用于根据维基百科人物传记数据集的事实表格生成人物传记的初始句子^[28]。

相比基于规则的方法,数据驱动的方法使得数据到文本生成更可能与领域无关,不需要专家参与,并且更容易优化,也更容易扩展。但是数据驱动的方法需要庞大的训练数据,而且训练数据的质量直接影响到训练模型的结果。

2 数据集

目前在特定领域已经公开了一些数据到文本生成的数据集,如表 1 所示。例如天气预报和体育比赛等领域,这些数据集基本都是由数据库记录 and 对应文本组成。天气预报领域的数据集有 SUMTIME-METEO^[29] 和 WEATHERGOV^[19],体育比赛领域的数据集有 ROBOCUP^[30]、NFL^[18]、ROTOWIRE 和 SBNATION^[4],航空领域的数据集有 ATIS^[31],人物传记领域的数据集有 WIKIBIO^[28]。数据集的使用方法在此不再赘述,详情请查阅相关文献。

表 1 可供下载的数据集

数据集名称	领域	下载地址
SUMTIME-METEO	天气预报	http://www.itri.brighton.ac.uk/home/Anja.Belz/Prodigy/
WEATHERGOV	天气预报	http://cs.stanford.edu/~pliang/papers/
ROBOCUP	体育比赛	http://www.cs.utexas.edu/~ml/clamp/sportscasting/
NFL	体育比赛	http://pages.cs.wisc.edu/~bsnyder/
ROTOWIRE SBNATION	体育比赛	https://github.com/harvardnlp/boxscore-data
ATIS	航空	http://www.ikonstas.net/index.php?page=resources
WIKIBIO	人物传记	https://github.com/DavidGranger/wikipedia-biography-dataset

3 评价方法

对于数据到文本生成来说,主要有两类评价:一类

是内在评价 (intrinsic evaluation), 通常和文本质量、输出正确性和可读性等问题相关; 另一类是外在评价 (extrinsic evaluation), 通常和任务完成有关, 即系统在做出决策时是否真正达到了目的。

3.1 内在评价方法

内在评价主要有两种方法, 一种依赖人类的判断即主观评价, 另一种基于语料。

人类判断的方法是通过专家根据某些标准评价系统输出。通常的标准有流畅性 (fluency) 或可读性 (readability), 即语篇的语言质量, 还有与输入有关的准确性 (accuracy)、充分性 (adequacy)、相关性 (relevance) 或正确性 (correctness), 反映了系统对内容的再现^[32]。

基于语料的评价方法是通过一些度量标准对比人类的输出和系统的输出。这种方式相对廉价, 常见的自动度量指标有 BLEU、NIST、ROUGE、F-measure 等。

3.2 外在评价方法

与内在评价不同, 外在评价衡量实现目标的有效性, 而有效性取决于应用领域和系统用途。通常基于问卷调查或者自我报告的研究可以解决外在评价, 但许多情况下评价需要依赖一些性能的客观衡量标准。外在评价又分为用户任务成功性度量 (user task success measure) 和系统目的成功性度量 (system purpose success measure)^[33]。

用户任务成功性度量衡量的是任何与用户从系统输出获得的有关的东西, 比如决策和理解准确性等。例如 2009 年 Gatt 等的 BabyTalk^[9] 使用了这种评价方法, 给用户展示两个输出, 用户做出决策, 以此来衡量哪个输出在决策中更有效。

系统目的成功性度量衡量一个系统是否能满足最初的目的。Reiter 等在 1999 年设计的 STOP 系统^[34] 为了帮助人们戒烟而生成简短的戒烟信, 使用这种评价方法来确定系统目的是否达到, 即用户是否戒烟。

外在评价对于判断一个数据到文本生成系统是否成功或者用户能否得到想要的东西来说非常重要, 也更有说服力。但这种评价方式在时间和费用上花费得更多, 而且依赖足够的用户基础, 并且必须有在现实中开展研究的可能性。

4 存在问题和发展方向

目前数据到文本生成存在一些问题, 需要在未来的研究中解决:

- (1) 数据集缺乏。可训练的数据集主要集中在天气和体育等几个专业领域, 数据集的建立需要人工收集数据甚至标注, 因此公开可用的数据集比较缺乏。
- (2) 生成文本短, 数据简单。数据集中生成的文

本长度较短, 用到的数据记录也较少, 因此在这些数据集上效果好的方法并不一定能满足复杂数据和生成长文本的需求。

(3) 评价方法不独立。适用于数据到文本生成的评价方法大多借鉴于机器翻译和文本摘要等领域, 没有单独完整的一套评价标准, 除了人类评价之外, 需要在自动度量标准上设计针对数据到文本生成的评价体系, 体现出内容的完整性、相关性、顺序结构以及表达性等等方面。

(4) 无法满足商业应用。用于商业的写作方法基本都是基于模板的方法, 成文较为固定, 虽然神经网络方法在实验阶段效果不错, 但在很多方面仍然不成熟, 暂时无法在商业中使用。

数据到文本生成虽然还存在很多问题, 但未来的发展方向仍然是不可限量的。比如, 结合视觉信息比单一使用图像或文字效果更好^[16]; 在领域之间或者语言之间转移学习方法^[35]; 研究处理不确定数据的方法, 大量数据是不确定的, 比如股票数据、天气数据或者网络数据等。近两年在数据到文本生成的研究中开始出现深度学习的方法并且获得了不错的效果, 相信随着神经网络的发展, 未来在该领域会有更多的研究者投入到使用神经网络的方法实现数据到文本生成的研究中来。

5 结束语

随着人工智能的发展, 数据到文本生成也越来越重要, 很多领域都在尝试使用机器代替部分人工, 完成自动文本的生成。国内外尤其是新闻行业, 在自动撰写新闻的尝试探索中竞争激烈, 但没有竞争就没有进步, 数据到文本生成需要各个领域的共同发展, 需要软硬件技术的不断推动, 只有存储和处理数据的能力越来越强, 神经网络方面的研究开展的更迅速, 数据到文本生成的研究和应用才能有更多的可能性。

参考文献:

[1] 李 苏. 机器新闻发展的市场进路及反思—以 Autamated Insights 公司为例[J]. 新闻界, 2015(18): 56-61.

[2] 周佳玥. 从 NLG 到机器新闻写作—机器新闻的发展与反思[J]. 今传媒, 2017, 25(10): 18-19.

[3] 金兼斌. 机器新闻写作: 一场正在发生的革命[J]. 新闻与写作, 2014(9): 30-35.

[4] WISEMAN S, SHIEBER S, RUSH A. Challenges in data-to-document generation[C]//Conference on empirical methods in natural language processing. [s. l.]: Association for Computational Linguistics, 2017: 2253-2263.

[5] SRIPADA S G, REITER E, HUNTER J, et al. A two-stage model for content determination[C]//Proceedings of the 8th

- European workshop on natural language generation – volume 8. Toulouse, France; Association for Computational Linguistics, 2001; 1–8.
- [6] HALLETT C, POWER R, SCOTT D. Summarisation and visualisation of e-health data repositories[C]//Proceedings of the UK e-science all-hands meeting. [s. l.]; National e-Science Centre, 2006.
- [7] TURNER R, SRIPADA S, REITER E, et al. Using spatial reference frames to generate grounded textual summaries of georeferenced data[C]//Proceedings of the fifth international natural language generation conference. Salt Fork, Ohio; Association for Computational Linguistics, 2008; 16–24.
- [8] THOMAS K E, SRIPADA S, NOORDZIJ M L. Atlas.txt; exploring linguistic grounding techniques for communicating spatial information to blind users[J]. Universal Access in the Information Society, 2012, 11(1): 85–98.
- [9] GATT A, PORTET F, REITER E, et al. From data to text in the neonatal intensive care unit; using NLG technology for decision support and information management[J]. AI Communications, 2009, 22(3): 153–186.
- [10] BLACK R, REDDINGTON J, REITER E, et al. Using NLG and sensors to support personal narrative for children with complex communication needs [C]//Proceedings of the NAACL HLT 2010 workshop on speech and language processing for assistive technologies. Los Angeles, California; Association for Computational Linguistics, 2010; 1–9.
- [11] TINTAREV N, REITER E, BLACK R, et al. Personal storytelling; using natural language generation for children with complex communication needs, in the wild... [J]. International Journal of Human-Computer Studies, 2016, 92–93; 1–16.
- [12] BANAEI H, AHMED M U, LOUTFI A. Towards NLG for physiological data monitoring with body area networks[C]//14th European workshop on natural language generation. [s. l.]; Association for Computational Linguistics, 2013; 193–197.
- [13] SCHNEIDER A, MORT A, MELLISH C, et al. MIME-NLG in pre-hospital care[C]//Proceedings of the 14th European workshop on natural language generation. [s. l.]; Association for Computational Linguistics, 2013; 152–156.
- [14] MANN W C, THOMPSON S A. Rhetorical structure theory; toward a functional theory of text organization[J]. Text-Interdisciplinary Journal for the Study of Discourse, 1988, 8(3): 243–281.
- [15] RAMOS-SOTO A, BUGARIN A J, BARRO S, et al. Linguistic descriptions for automatic generation of textual short-term weather forecasts on real prediction data [J]. IEEE Transactions on Fuzzy Systems, 2015, 23(1): 44–57.
- [16] GKATZIA D, LEMON O, RIESER V. Natural language generation enhances human decision-making with uncertain information[C]//Proceedings of the 54th annual meeting of the association for computational linguistics. [s. l.]; Association for Computational Linguistics, 2016; 264–268.
- [17] DUBOUE P A, MCKEOWN K R. Statistical acquisition of content selection rules for natural language generation[C]//Proceedings of the 2003 conference on empirical methods in natural language processing. [s. l.]; Association for Computational Linguistics, 2003; 121–128.
- [18] BARZILAY R, LAPATA M. Collective content selection for concept-to-text generation[C]//Proceedings of the conference on human language technology and empirical methods in natural language. Vancouver, British Columbia, Canada; Association for Computational Linguistics, 2005; 331–338.
- [19] LIANG P, JORDAN M I, KLEIN D. Learning semantic correspondences with less supervision[C]//Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP; volume 1. Suntec, Singapore; Association for Computational Linguistics, 2009; 91–99.
- [20] ANGELI G, LIANG P, KLEIN D. A simple domain-independent probabilistic approach to generation[C]//Proceedings of the 2010 conference on empirical methods in natural language processing. Cambridge, Massachusetts; Association for Computational Linguistics, 2010; 502–512.
- [21] KONSTAS I, LAPATA M. Unsupervised concept-to-text generation with hypergraphs[C]//Conference of the North American chapter of the association for computational linguistics; human language technologies. Montreal, Canada; Association for Computational Linguistics, 2012; 752–761.
- [22] KONDADADI R, HOWALD B, SCHILDER F. A statistical NLG framework for aggregated planning and realization [C]//Proceedings of the 51st annual meeting of the association for computational linguistics. [s. l.]; Association for Computational Linguistics, 2013; 1406–1415.
- [23] SOWDABOINA P K, CHAKRABORTI S, SRIPADA S. Learning to summarize time series data[C]//Proceedings of the 15th international conference on computational linguistics and intelligent text processing – volume 8403. Kathmandu, Nepal; Springer-Verlag, 2014; 515–528.
- [24] GKATZIA D, HASTIE H, LEMON O. Comparing multi-label classification with reinforcement learning for summarisation of time-series data[C]//Proceedings of the 52nd annual meeting of the association for computational linguistics. [s. l.]; [s. n.], 2014; 1231–1240.
- [25] GKATZIA D, HASTIE H, JANARTHANAM S, et al. Generating student feedback from time-series data using reinforcement learning[C]//Proceedings of the 14th European workshop on natural language generation. [s. l.]; Association for Computational Linguistics, 2013; 115–124.
- [26] MAHAPATRA J, NASKAR S K, BANDYOPADHYAY S. Statistical natural language generation from tabular non-text-

由实验结果可知,在数据量较小时,由于 Spark 的启动需要消耗大量资源以及时间,因而无法体现并行化算法在时间效率方面的优势,但随着数据量的增加,其时间效率明显提升。

5 结束语

设计了一种 Item-Based 协同过滤算法在 Spark 集群中的并行化方案,并通过基于 MovieLens 数据集的实验结果证明,在应对大规模数据处理时,基于 Spark 的并行化 Item-Based 协同过滤算法,不仅可以保证评分的准确性,而且算法执行速度更快,可以提高推荐系统的时效性。

参考文献:

[1] GOLDBERG D, NICHOLS D, OKI B M, et al. Using collaborative filtering to weave an information tapestry[J]. Communications of the ACM, 1992, 35(12): 61-70.

[2] 冷亚军, 陆青, 梁昌勇. 协同过滤推荐技术综述[J]. 模式识别与人工智能, 2014, 27(8): 720-734.

[3] BELLOGÍN A, CASTELLS P, CANTADOR I. Neighbor selection and weighting in user-based collaborative filtering[J]. ACM Transactions on the Web, 2014, 8(2): 1-30.

[4] 李涛, 王建东, 叶飞跃, 等. 一种基于用户聚类的协同过滤推荐算法[J]. 系统工程与电子技术, 2007, 29(7): 1178-1182.

[5] 何哲. 基于用户聚类的推荐算法研究[J]. 科技创业月刊, 2017, 30(10): 135-136.

[6] 邓爱林, 左子叶, 朱扬勇. 基于项目聚类的协同过滤推荐算法[J]. 小型微型计算机系统, 2004, 25(9): 1665-1670.

[7] KIM B M, LI Q, PARK C S, et al. A new approach for combining content-based and collaborative filters[J]. Journal of Intelligent Information System, 2006, 27: 79-91.

[8] 邓爱林, 朱扬勇, 施伯乐. 基于项目评分预测的协同过滤推荐算法[J]. 软件学报, 2003, 14(9): 1621-1628.

[9] TIAN X H. PCIB: a new algorithm for item-based collaborative filtering recommendations[C]//Proceedings of 2014 international conference on artificial intelligence and industrial application. [s. l.]: Advanced Science and Industry Research Center, 2014: 9.

[10] 杨志伟. 基于 Spark 平台推荐系统研究[D]. 合肥: 中国科学技术大学, 2015.

[11] 赵娟, 程国钟. 基于 Hadoop、Storm、Samza、Spark 及 Flink 大数据处理框架的比较研究[J]. 信息系统工程, 2017(6): 117.

[12] 唐振坤. 基于 Spark 的机器学习平台设计与实现[D]. 厦门: 厦门大学, 2014.

[13] 徐新瑞, 孟彩霞, 周雯, 等. 一种基于 Spark 时效化协同过滤推荐算法[J]. 计算机技术与发展, 2015, 25(6): 48-55.

[14] 李成, 冯青青. 推荐系统准确度衡量方案—引入权重概念[C]//工业设计研究. 出版地不详: 出版者不详, 2017: 269-275.

(上接第 84 页)

tual data[C]//Proceedings of the 9th international natural language generation conference. [s. l.]: Association for Computational Linguistics, 2016: 143-152.

[27] MEI H, BANSAL M, WALTER M R. What to talk about and how? selective generation using LSTMs with coarse-to-fine alignment[C]//Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics; human language technologies. [s. l.]: Association for Computational Linguistics, 2016: 720-730.

[28] LEBRET R, GRANGIER D, AULI M. Neural text generation from structured data with application to the biography domain[C]//Proceedings of the 2016 conference on empirical methods in natural language processing. [s. l.]: [s. n.], 2016: 1203-1213.

[29] REITER E, SRIPADA S, HUNTER J, et al. Choosing words in computer-generated weather forecasts[J]. Artificial Intelligence, 2005, 167(1-2): 137-169.

[30] CHEN D L, MOONEY R J. Learning to sportscast: a test of grounded language acquisition[C]//Proceedings of the 25th

international conference on machine learning. Helsinki, Finland: ACM, 2008: 128-135.

[31] DAHL D A, BATES M, BROWN M, et al. Expanding the scope of the ATIS task; the ATIS-3 corpus[C]//Workshop on human language technology. [s. l.]: [s. n.], 1994: 43-48.

[32] GATT A, KRAHMER E. Survey of the state of the art in natural language generation: core tasks, applications and evaluation[J]. Journal of Artificial Intelligence Research, 2018, 61: 65-170.

[33] GKATZIA D, MAHAMOOD S. A Snapshot of NLG evaluation practices 2005-2014[C]//Proceedings of the 15th European workshop on natural language generation. [s. l.]: Association for Computational Linguistics, 2015: 57-60.

[34] REITER E, ROBERTSON R, OSMAN L. Types of knowledge required to personalise smoking cessation letters[C]//Joint European conference on artificial intelligence in medicine and medical decision making. [s. l.]: [s. n.], 1999: 389-399.

[35] PAN S J, YANG Q. A survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345-1359.