利用人工智能实现自主编辑系统与计算调查

Ahmed Banafea
Betron Labs LLC

2026年3月17日

arXiv:2603.13232v1 [cs.DL] 23 Jan 2026

摘要

理解新闻不仅仅是报道个别事实。它还要求判断哪些报道故事至关重要、不同来源的信息如何相互关联，以及随着时间推移对报道进行整体考察后会得出什么结论。随着公共新闻记录的规模不断扩大、连续性增强且愈发碎片化，这些工作越来越多地需要在系统级方法发挥效用的规模上开展。

本文介绍一类能够自动执行两项核心职能的人工智能系统：自主编辑组织和计算调查。这些系统利用大语言模型进行语义解释，以基于嵌入的表示来衡量相似性并进行对齐，并通过结构化存储维护持久的编辑状态。这些组件协同运作，能够将大量报道整理成长存的报道故事结构，并通过跨来源、跨时间的系统比较加以考察。

本文提出的框架直接源于 World Pulse Now（WPN News）[1], 的设计与运行；这是一个目前已投入生产运行的代理型编辑系统。虽然 WPN News 对外定位为自主编辑组织，但其底层架构支持更广泛的能力：利用人工智能，通过大规模比较与推理，对公共记录展开调查。

通过将报道故事视为持久的计算对象，将编辑组织与调查分析相分离，并以确定性方式协调人工智能组件，该系统将信息规模从一种限制转变为洞察来源。

1 引言

当代公共新闻的规模已经超出仅凭人工编辑流程所能连贯解读的范围。重大全球事件通常会在不同媒体、地区和政治背景下，于短时间内产生数百篇乃至数千篇文章。尽管这种丰富性提升了信息获取的便利性，却也带来了协调问题：如何理解大量报道共同描绘的不断演变的事件。

理解新闻的意义，不仅仅是阅读单篇文章。这一过程还包括识别哪些报道指向同一潜在事件，理解不同来源的叙述有何差异，跟踪 Narratives 如何随时间变化，以及判断新的进展何时会实质性地改变整体面貌。传统上，这些任务由人工编辑和调查人员承担，同时面临显著的注意力与时间约束。

随着报道量和连续性不断增加，这些活动开始呈现出计算问题的特征。信息检索和信息抽取领域的既有研究已将大规模

[1]https://www.worldpulsenow.com

文本流视为事件导向系统，需要跨文档、跨时间持续进行检测、监控和比较 [1, 2]。人工智能领域的最新进展——尤其是大语言模型和表征学习——使我们能够通过系统设计而非依赖偶发的人类判断来应对这些挑战。

本文探讨了如何将人工智能的两项核心能力直接作为系统功能加以实现：自主编辑组织和计算调查。与逐案协助人类不同，本文所述系统能够持续、大规模地阅读、比较和汇总报道，同时处理数千篇报道故事。编辑结构和调查洞见由系统自身的运行过程涌现。

本文提出的理念源于 World Pulse Now（WPN News）的设计与部署。这是一个持续组织和分析公共新闻记录的智能体系统。尽管 WPN News 对外主要聚焦于自主编辑组织，但系统的构建揭示了一个更广泛的洞见：在拥有足够独立报道的前提下，人工智能不仅能够将新闻组织成一致且不断演进的报道结构，还能够通过结构化比较和推理，对公共记录开展调查。

2 编辑系统与规模化调查

新闻生产传统上围绕不同的人类角色组织。记者负责收集信息并描述事件。编辑决定如何对报道故事进行归类、排序和解释。调查人员跨越不同来源和时间进行审视，以识别模式、时间线及更广泛的影响。这种劳动分工在小规模情境下运作良好。

然而，在大规模场景下，它会遭遇人类能力的限制。没有哪位编辑能够在较长时期内持续跟踪数千篇相关文章。没有哪个调查团队能够系统地比较数十个来源针对某一发展中事件所提出的每一项主张。随着报道量增长，瓶颈不再是获取信息，而是综合信息的能力。

许多调查洞见源于综合，而非发现新的事实。当多个来源提供了零散、不断演变或相互冲突的叙述时，新的理解会通过比较产生。此前关于流式文本语料库中的事件检测、跟踪和故事演变的计算研究，已经展示了如何在大规模场景下对这类动态进行建模[3, 5, 4]。

人工智能尤其适合这一场景。人工智能系统能够持续阅读，在大型语料库中维护结构化记忆，并在较长时间跨度内应用一致的比较逻辑。因此，编辑组织和调查工作都会自然地转向持久运行的计算过程，而不再依赖间歇性的人工作预干预。

这种转变使我们能够以不同方式思考编辑工作和调查。它们不再被视为间歇性执行的手动任务，而是可以被设计为系统级操作：持续运行、在大规模场景下组织、比较并解读报道的过程。

3 自主编辑系统

编辑工作由一系列具体操作组成，这些操作可以独立于新闻编辑室岗位进行描述。这些操作包括识别哪些文章描述的是同一事件、确定信息是否为新信息、跟踪事件如何随时间变化、评估相对重要性，以及生成反映当前事态的解释。

自主编辑系统利用人工智能直接执行这些操作。这类系统会摄取大量报道，解读文章内容，并将相关内容归类

为有序的事件结构。随着新文章不断到来，系统会更新对每个事件的表征，并据此调整编辑输出。

自主编辑系统的一个显著特征是，它们将报道故事视为持久的有状态结构，而非转瞬即逝的文章集合。文章作为输入，而报道故事则随着时间推移不断积累背景、历史和结构。这种方法与早期的系统设计相契合：在这些设计中，文档被视为不断演化的叙事或事件模型的输入 [3, 4]。

编辑成果由系统本身生成。人类定义系统目标、约束和运行参数，但不会在发布时选择具体的标题、摘要或排名。编辑组织由此成为直接嵌入系统架构的持续计算过程。

自主编辑系统不同于个性化或推荐系统，后者围绕用户行为进行优化。自主编辑系统旨在支持公共记录中的叙事连贯性、连续性和解释清晰度。

图1展示了整体系统架构，说明文章如何从摄取开始，经过丰富化和嵌入，最终进入支持编辑组织的持久化报道故事结构。

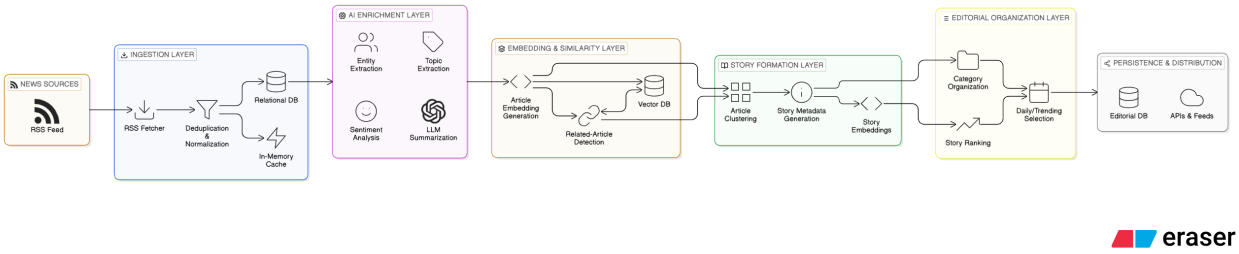

图1：自主编辑系统的端到端架构，从文章摄取到编辑输出。

4 计算调查

调查可以理解为将多条信息放在一起审视，从而发现其中浮现内容的过程。计算调查借助人工智能，将这一理念大规模应用。

该系统并不收集新的证据，而是考察现有报道之间的关系。它比较不同来源如何描述同一事件，追踪主张如何随时间演变，并识别趋同、分歧、延迟确认和叙事转变等模式。此前的研究已表明，这种纵向跨来源方法有助于理解故事演变和事件动态 [5, 6]。

当相互独立的来源对某项主张达成一致时，解读置信度会增强。当不同说法出现分歧或发生变化时，转变便会显现。这些信号来自对大量报道的系统比较，而非孤立地阅读单篇文章。

人工智能通过维护主张的结构化表示，并在不同来源和时间之间对其进行监测，使这一过程得以实现。调查由此成为一种推理活动，理解从结构、重复和变化中涌现。

计算调查与自主编辑系统并行运作。编辑组织总结当前已知的信息，而计算调查则探究对更广泛的公共记录进行整体审视后所揭示的内容。

5 作为计算基底的公共记录

对于自主编辑系统和计算调查，公共新闻记录被视为结构化输入，而非孤立文档的集合。每篇文章都包含主张、参考资料、实体、来源和时间戳等元素，这些元素可以被明确提取并加以表示 [2]。

汇总之后，这些元素构成了丰富的计算基底。主张在不同来源之间相互关联，报道故事随着时间推移而延续，随着报道不断累积，叙事轨迹也变得清晰可见。以往关于叙事可视化和时间分析的研究强调，持久化表示对于理解信息如何演变至关重要 [7]。

人工智能系统通过识别主张之间的关系，并随着新报道的出现持续更新这些关系，在这一计算基底上运行。来源独立性会增强信号，而分歧则凸显不确定性或转变。时间深度使人们能够在更长的时间跨度内考察模式，而非仅仅观察某个单一时刻。

由于报道故事会持续存在并不断累积背景信息，每增加一篇文章，都会提升系统的信息价值。图2展示了文章如何被组织成持久化报道故事结构，在不同来源和时间之间逐步构建背景信息。

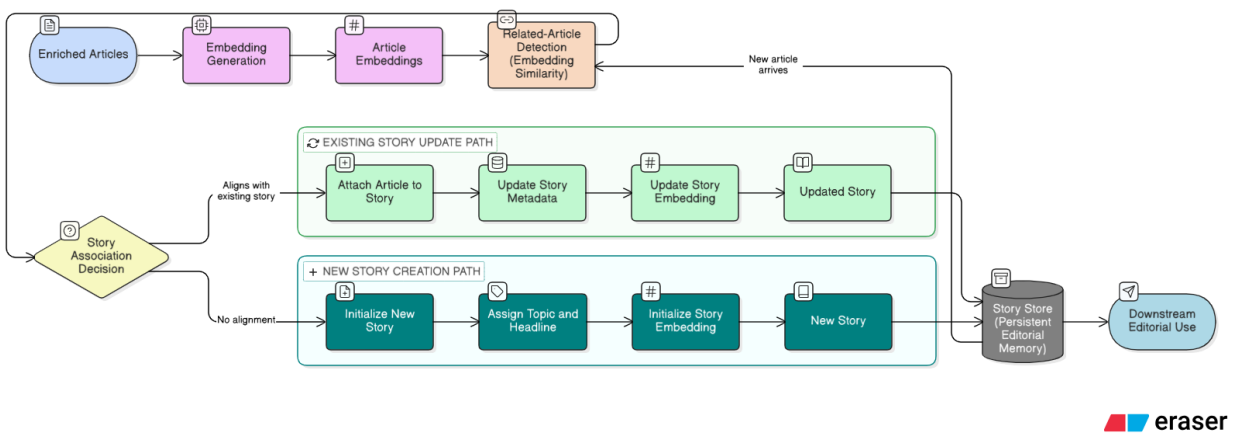

图2：基于嵌入的故事形成与持久化编辑记忆。新收到的文章会与现有报道故事关联，或开启新的报道故事，使报道故事能够随时间演变。

6 系统架构

该系统架构通过一条持续处理新到报道的结构化管线，实现自主编辑组织与计算调查。它将报道故事视为持久化状态，将编辑组织与调查分析分开，并以确定性方式编排人工智能组件。

图1展示了完整的系统架构，而图3聚焦于文章级处理流水线，涵盖从摄取到丰富化和持久化的全过程。这些图共同说明了单篇文章如何转化为结构化表示，并整合进长期存在的编辑结构中。

在这一架构中，人工智能指的是多种技术的组合：用于语义解释和比较的大语言模型、用于相似性计算和聚类的基于嵌入的表示，以及用于维护编辑状态的结构化存储。这些组件通过一条受控流水线进行协调；在这条流水线中，每个模型都执行特定且明确定义的功能，而非自主运行。

这一设计强调长期的一致性与连续性。关于表征学习中的语义稳定性和知识保留的近期研究，凸显了在长期运行的系统中保持解释连贯性的重要性 [8, 9]。本文提出的架构体现了这一原则：随着新的报道不断到来，编辑结构逐步演进，而非彼此孤立地重新计算。

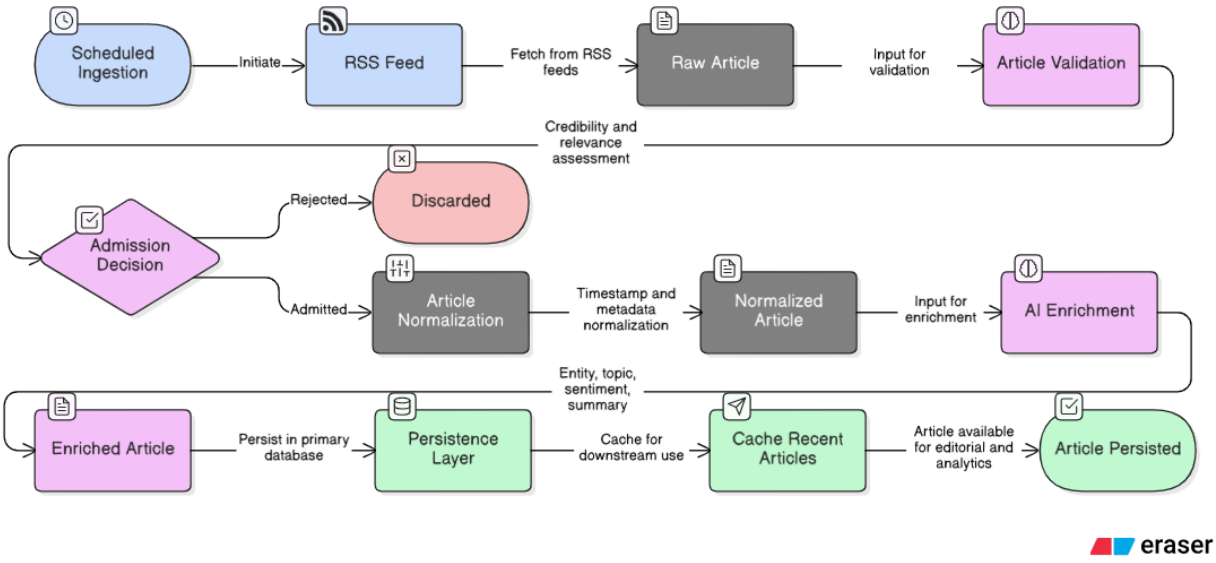

图3：文章摄取与基于人工智能的增强流水线。文章经过验证、规范化和丰富化后，通过定时后台处理予以持久化存储。

7 回顾性案例研究

本节介绍一个取材于实时系统运行的回顾性案例研究。该案例展示了随着文章不断到来，系统如何将报道组织成持久的编辑结构，以及编辑组织如何通过持续处理而非人工干预逐步形成。

此示例聚焦于故事形成和累积，这是更广泛的系统架构中一个可见的组成部分。

7.1 观察窗口与文章语料库

观察窗口为期三天，从2026年1月18日至1月20日。在此期间，系统将来自7家独立新闻机构的12篇文章归入同一个故事簇。

文章的出现遵循了许多主题中常见的模式：早期相关文章零星出现，随着主题受到关注，报道浪潮逐渐变得密集。随着报道量和来源多样性不断增加，这一进程使我们得以观察编辑结构如何随时间发展。

Date	文章	独立来源	编辑状态
2026年1月18日	3	3	已识别相关报道
2026年1月19日	3	3	持续累积
2026年1月20日	7	5	形成持续报道主题

表1：案例研究故事的文章与来源累积。

7.2 故事形成与编辑持久化

随着文章不断到来，系统会评估语义相似性、时间邻近性和跨来源对齐，以确定新收到的报道是否应与现有结构相关联。

当累积的报道在密度和连贯性上超过某个阈值时，系统便会创建一个持久化的故事对象，并追溯性地关联此前相关的文章。从那时起，该故事便作为一个稳定的编辑单元发挥作用，能够不断积累更多背景信息。

这种行为体现了系统的一项通用设计原则：只有在信息密度足以支持连续性时，编辑结构才会逐渐形成。报道会随时间延续，并作为累积报道的容器，而非与单次发布时刻绑定。

7.3 跨来源聚合与编辑记忆

随着报道不断累积，来自多个独立来源的文章被整合到一个共同的编辑结构中。每篇文章都会为故事的内部状态补充信息，使其能够在不同日期和出版方之间保持连续性。

随着报道增多，故事表征变得更加稳定。不同来源反复出现的要素强化了共同语境，而孤立细节的影响则相对减弱。这一累积过程使系统能够保持编辑记忆，同时持续响应新信息。

8 相关工作

关于事件检测、抽取和跟踪的研究探讨了如何将大型文本流组织成有意义的结构 [1, 2]。后续工作将这些理念拓展到流式处理和社交数据领域，对故事演变和主题动态随时间的变化进行建模 [3, 4]。

其他研究探索了叙事结构和时间可视化，以理解不断演变的信息 [7, 5]。较新的系统则关注增量式和流式处理场景中的知识积累与语义稳定性 [8, 9]。

本文将这些脉络整合为统一的系统架构，使自主编辑组织和计算调查成为系统持续具备的能力。

9 结论

本文介绍了一种面向系统的方法，利用人工智能实现自主编辑系统和计算调查。通过组织和分析大量独立报道，该系统将信息规模从一种限制转化为洞见来源。

本文所述框架直接源自 World Pulse Now（WPN News）的设计与运营，并可推广至新闻以外的其他信息丰富领域。持久化报道故事结构、编辑记忆和结构化比较，为可扩展的理解奠定了基础，随着公共信息生态系统不断扩展，这一基础将愈发重要。

参考资料

[1] Y. Chen 等人，《文本事件检测与识别的历史与理论》。*IEEE Access*，2020年。[2] W. Xiang 和 B. Wang，《文本事件抽取综述》。*IEEE Access*，2019年。[3] A. Ahmed 等人，《流式新闻的统一分析》。*WWW* 论文集，2011年。[4] H. Peng 等人，《流式社交事件检测与演变发现》。*ACM Transactions*，2021年。[5] S. Rose 等人，《描述动态信息流中的故事演变》。*IEEE Symposium*，2009年。[6] I. Mele 等人，《异构新闻流中的事件挖掘与时效性分析》。信息处理与管理，2019年。[7] D. Fisher 等人，《Narratives：用于追踪叙事事件的可视化工具》。*IEEE Symposium*，2008年。[8] Y. Cao 等人，《知识保持型增量社交事件检测》。*The Web Conference*，2021年。

[9] M. Lozano 等人：《大语言模型中的语义深度再分配》，2024年。