NewsCube: A Multi-Dimensional HyperCube Representation for Interactive Exploration of Enterprise News Corpora

Sangameshwar Patil, Ajay Phade, Alok Kumar
TCS Research
firstname.lastname@tcs.com

Abstract

In this paper, we demonstrate an intelligent, personalized user interface for enterprise users who are inundated with humongous amount of business relevant news data. These users need interfaces that enable fast and interactive exploration of large and evolving corpora of news articles. Existing user interfaces either focus on relative frequency or co-occurrences of keywords (e.g., word clouds, clustering, Topic modeling) or organize the news using timeline representation and its recent variants such as Story Forest. All these representations ignore that there are different latent aspects or dimensions about a news article which are of particular interest from an enterprise user's perspective. In this paper, we propose NewsCube, a multi-dimensional hypercube representation that overcomes the limitations of existing interfaces to explore news corpora. Comparison of NewsCube based user-interface with other UI representations by experienced business users and senior management has confirmed the ease of navigation, high utility and on-demand focused interactivity of the proposed NewsCube representation. The accompanying video gives an overview of the NewsCube.

Introduction

In many enterprises, it is important to keep track of the recent business as well as domain-specific news while maintaining their contextual connections with the earlier events described in the news articles in the past. For example, the leadership team as well as the sales and marketing executives need to continually keep track of the evolving business and technology landscape, get insights into market trends, policy and regulatory changes. These activities are critical for their work as they need to identify growth and transformation opportunities and take strategic decisions periodically. They need tools that can enable them to focus on relevant subsets of the news related to specific entities (e.g., a particular organization, technology or industry sector etc.) on-demand. For this purpose, user interfaces that enable fast and interactive exploration of large and evolving corpora of news articles are required.

Multiple interfaces backed by different analytics-enabled representations have been proposed towards this goal. Word cloud has been a traditionally popular technique based on the relative frequency of keywords. Document clustering and Latent Dirichlet Allocation (LDA) based Topic Modeling exploit word co-occurrence statistics to identify similarities among the documents in a corpus. A complementary line of work focuses on the temporal relations among the documents. These relations are easy to extract for a news corpus given the natural ordering due to the publication date. Timeline based summarization of news articles is one such popular representation and has received active attention from the research community (Ghalandari and Ifrim 2020; Chen et al. 2023). Story Forest, a more sophisticated variation of timeline has been proposed (Liu et al. 2020) to track a particular story, phenomenon or event spanning over multiple news articles.

However, all these approaches have limited utility for the comprehensive, multi-faceted, interactive and on-demand analysis as required by enterprise users mentioned above. All these representations ignore that there are different aspects or dimensions of interests about a news article from the reader or consumer's perspective. This is especially evident in case of business news which is relevant for the enterprise users mentioned earlier. A business news can be associated with one or more *Industry sectors* (e.g., manufacturing, telecom, healthcare, etc.), *Geographies* (e.g., Europe, America, Asia), *Technologies* (e.g., AI, Cloud computing, Blockchain, 5G) and so on. But none of the above mentioned representations make use of these latent aspects while providing an interface for exploring a news corpus. There has been work on OnLine Analytical Processing (OLAP) cube (Gray et al. 1997) that can preprocess the numeric and categorical data in relational databases and organise it in OLAP data structures to enable fast and interactive business intelligence. However, OLAP-like interface is not available for unstructured text in news corpora.

The key contribution of this paper is to fill this gap by proposing NewsCube, a multi-dimensional hyper-cube representation that overcomes these limitations. We first give an overview of the NewsCube with a brief illustrative example, followed by comparative analysis. Finally, we make concluding remarks and outline future work.

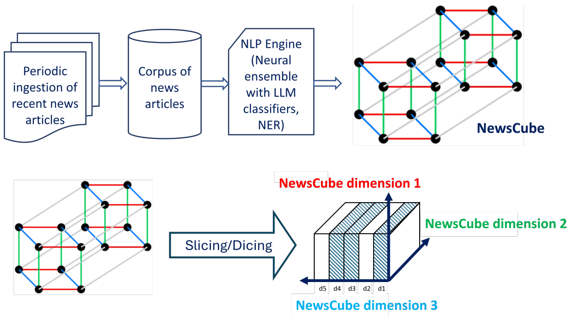

Figure 1: Preprocessing pipeline to construct NewsCube and the Slicing-Dicing operations that allow user to focus on subset of interest

NewsType vs Org in (AI tech.)	OpenAI	Google	Microsoft	Meta
Product Launches/ Offerings	OpenAI Launches GPT-4 API For Everyone	Google's AI-assisted note-taking app gets limited launch as NotebookLM	Microsoft brings new AI-powered shopping tools to Bing and Edge	Meta Introduces Generative AI Model 'CM3leon' For Text, Images
Research & Innovation	OpenAI is forming a new team to bring 'superintelligent' AI under control	Google's Med-PaLM AI model: Revolutionising healthcare with rapid medical knowledge retrieval	Moody's and Microsoft Develop Collaboration Solutions Powered by Generative AI	Meta AI Develops Visual Affordance Model Using Internet Videos of Human Behavior
Legal	OpenAI faces lawsuits over alleged copyright infringements	Google faces lawsuit alleging data scraping and copyright violations for AI development	OpenAI & Microsoft have created an 'AI arms race': Lawyer suing the companies	Texas attorney general files lawsuit against Meta over Facebook's facial recognition technology

Figure 2: Example: A subset of articles obtained by Slicing-Dicing operations to focus on *NewsType* vs. *Organization* dimensions of the NewsCube for AI *Technology* new.

NewsCube: A Multi-Dimensional Hypercube-based Representation

Intuition behind the NewsCube is to embellish the unstructured text in the news articles with metadata about different latent dimensions which are of interest to the enterprise user groups. We first preprocess the news corpus and remove duplicate articles using near duplicate detection (Kumar, Tumre, and Patil 2025; Tumre, Patil, and Kumar 2025). This helps to reduce redundant information while creating NewsCube. We then assign metadata using natural language processing (NLP) techniques that make use of the classifiers based on Neural Ensembles (Ramrakhiyani et al. 2023) as well as Named Entity Recognition (NER). This metadata can be assigned in the form of categories or tags corresponding to each dimension of interest as well as named entities mentioned in the news text (e.g., names of organizations, locations mentioned in the news). The dimensions and entity types of interest are customizable based on the requirements of an enterprise and can be given as input to NLP engine. Fig. 1 summarizes this preprocessing pipeline for NewsCube.

As an illustration, we give sample *dimensions* followed by the categories in each dimension for a NewsCube – *(Dim.1) NewsType:* Product Launch, Research-Innovation, Merger-Acquisition, Legal, . . . ; *(Dim.2) Industry Sector, i.e., Domain:* : Telecom, Healthcare, Manufacturing, . . . ; *(Dim.3) Technology:* AI, Cloud, Blockchain, 5G, . . . ; *(Dim.4) Geography* (hierarchical dimension: names of City-State-Country); *(Dim.5) Organization:* (specific named entities of interest); *(Dim.6) Time* (hierarchical dimension: day-week-month-quarter-year). These dimensions can be looked at as basis of a six-dimensional vector space. The metadata tagged to the articles in news corpus by NLP engine (Figure 1) helps to map the articles to this hyperspace. Hence, the NewsCube corresponds to a six-dimensional hypercube in this space.

The NewsCube facilitates the OLAP-like systematic slicing-dicing operations which were not possible with the raw news corpus. This enables the enterprise users to focus on a particular subset of news articles that are of interest. For instance, Fig. 2 shows a subset of news articles sliced and diced along the *NewsType* and *Organization* dimensions when only a subset of articles related to AI *Technology* are under focus. The video accompanying this paper shows more examples and an overview of NewsCube.

Comparative Analysis

We carried out qualitative analysis of NewsCube-based user interface with Timeline and Story Forest based interfaces. We showed the user interfaces based on NewsCube as well as TimeLine and Story Forest to multiple (10+) experienced industry experts and senior business leaders. The unanimous feedback received from them clearly indicated that NewsCube's slicing and dicing based interactivity was far more useful than other two representations. Timeline representation focuses only on the temporal aspect among the news and merges unrelated news articles just because their temporal proximity. Story Forest algorithm tends to generate too many short trees. The experienced professionals found it difficult to keep track of the trees and navigate the corpus using them. Compared to TimeLine, Story Forest and other representations, NewsCube enables easy, on-demand navigation through vast news corpora and allows to zero-in on the subset of user-specific interesting news. Further, the UI based on NewsCube utilizes the available display area more efficiently and on average, packs more relevant content. Further, as shown in the video, the UI based on Newscube enables heat-map based color-coding to provide additional visual cues and personalize the NewsCube interface based on the user's areas of interests.

Conclusion

In this paper, we proposed NewsCube, a multi-dimensional hypercube representation built by extracting user relevant metadata from news corpora. NewsCube allows on-demand interactive exploration of focused subsets of news that are of interest to the enterprise users. It enables an OLAP-like slicing and dicing for fast and responsive interface. Comparative analysis showed that NewsCube is able to overcome limitations of existing representations like timeline and its variants for news corpora exploration.

References

Chen, X.; Li, M.; Gao, S.; Chan, Z.; Zhao, D.; Gao, X.; Zhang, X.; and Yan, R. 2023. Follow the timeline! generating an abstractive and extractive timeline summary in chronological order. *ACM Trans. Inf. Syst.* 41(1).

Ghalandari, D. G., and Ifrim, G. 2020. Examining the state-of-the-art in news timeline summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Gray, J.; Chaudhuri, S.; Bosworth, A.; Layman, A.; Reichart, D.; Venkatrao, M.; Pellow, F.; and Pirahesh, H. 1997. Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals. *Data mining and knowledge discovery* 1:29–53.

Kumar, A.; Tumre, S.; and Patil, S. 2025. Benchmarking near-duplicate detection in the era of pay-walled news. In *Companion Proceedings of the ACM Web Conference 2025*, WWW '25. Association for Computing Machinery.

Liu, B.; Han, F. X.; Niu, D.; Kong, L.; Lai, K.; and Xu, Y. 2020. Story forest: Extracting events and telling stories from breaking news. *ACM Trans. on KDD* 14(3).

Ramrakhiyani, N.; Patil, S.; A., K.; and G., P. 2023. Neural ensemble for organizing business news-streams for corporate libraries. In *Proceedings of the ACM/IEEE Joint Conf. on Digital Libraries (JCDL)*.

Tumre, S.; Patil, S.; and Kumar, A. 2025. Improved near-duplicate detection for aggregated and paywalled news-feeds. In *2025 Annual Conference of the North American Chapter of the Association for Computational Linguistics Industry Track*.