Chrono Clustering: A Novel Methodology for Dynamic Topic Trend Analysis

Qiaomu Li
School of Data Science
Kennesaw State University
Marietta, GA 30060, USA
qli12@students.kennesaw.edu

Ying Xie
Department of Information Technology
Kennesaw State University
Marietta, GA 30060, USA
yxie2@kennesaw.edu

Shaoen Wu
Department of Information Technology
Kennesaw State University
Marietta, GA 30060, USA
swu10@kennesaw.edu

Abstract— **We develop a new framework, Chrono Clustering, to uncover and portray the progression of topics over time, where dual clustering, enhanced keyword extraction and hypergraph visualization are integrated. With K-Means on embedding space for base topic discovery and one-iteration K-Medoids for aligning base topics to other timeframes, our novel method could effectively quantify temporal topic shifts. Evaluated on AI conference datasets, Chrono Clustering boosts detection of trending topics that match real-world advances. It generates novel Chrono Graph to intuitively show dynamic topic progressions, demonstrating promising values for temporal topic modeling**

I. Introduction

Topic modeling is an unsupervised learning technique that can extract hidden thematic structures from large collections of unstructured text data (Blei, 2012). It has become a vital tool for organizing, summarizing, and gleaning insights from massive corpora. Popular topic modeling methods include Latent Dirichlet Allocation (LDA), which represents topics as multinomial distributions over words (Blei et al., 2003). LDA posits that each document exhibits multiple topics to different degrees. Another common technique is Latent Semantic Analysis (LSA), which applies singular value decomposition to reduce the dimensionality of the word-document matrix and uncover relationships between terms and documents (Landauer et al., 1998).

However, these conventional topic modeling approaches fundamentally assume that topics are static distributions. They fail to capture how the prevalence and composition of topics evolve over time (Schofield & Mimno, 2016). There are difficulties for analyzing temporal collections such as scientific literature, where the popularity of research areas rises and falls. For instance, when examining research papers across years, the focus of academic fields changes as new paradigms emerge while older ones fade. Traditional statistical topic modeling does not account for such chronological dynamism.

This paper presents Chrono Clustering, a novel method designed to analyze and visualize the temporal changes of topics within textual data. Our motivation lies in sharpening the detection and depiction of these evolutions, enhancing the depth with which we can observe and understand topic dynamics to gain actionable insights. From this method, researchers can learn innovative techniques for capturing and visualizing the fluidity of topics over time, applying this knowledge to domains where trend analysis is essential.

As illustrated in Fig.1, our approach first applies K-Means clustering on vector embeddings from the base dataset to establish initial topic clusters. These clusters represent salient topics in the current timeframe and serve as anchors. We then align other timeframe datasets to these topics using one-iteration K-Medoids. This enables quantifying topic evolution by comparing temporal centroids. We also recompute topic popularities locally to discern trends amidst shifts. By setting thresholds, we can spotlight meaningful temporal changes.

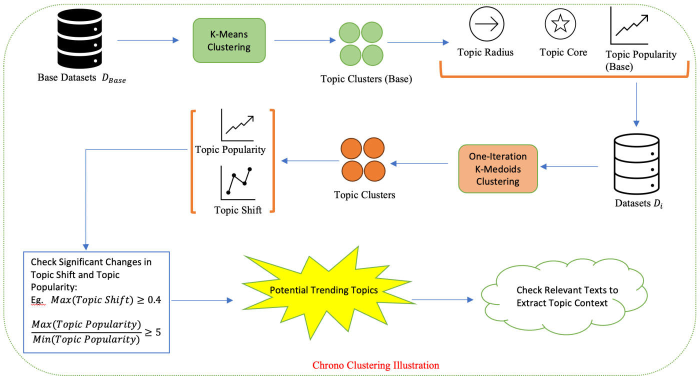
Base Datasets D_{Base}
K-Means Clustering
Topic Clusters (Base)
Topic Radius
Topic Core
Topic Popularity (Base)
Datasets D_i
One-Iteration K-Medoids Clustering
Topic Clusters
Topic Popularity
Topic Shift
Check Significant Changes in Topic Shift and Topic Popularity:
Eg. $Max(Topic\ Shift) \geq 0.4$
$\frac{Max(Topic\ Popularity)}{Min(Topic\ Popularity)} \geq 5$
Potential Trending Topics
Check Relevant Texts to Extract Topic Context
Chrono Clustering Illustration

Fig. 1. Chrono Clustering framework architecture

A unique aspect of Chrono Clustering is the tailored keyword extraction technique it employs for trending topics. As depicted in Fig 2, it extracts semantically representative terms from pertinent documents across different time periods. These keywords are then visualized with hypergraph showing the progression of the topic via both consistent and diverging terms.

Overall, our proposed framework combines clustering, temporal realignment, evolutionary analysis, keyword extraction and graphical visualization to provide insights into the dynamics of topics over time.

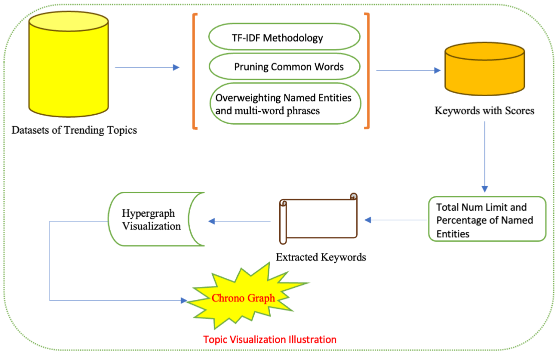
Datasets of Trending Topics
TF-IDF Methodology
Pruning Common Words
Overweighting Named Entities and multi-word phrases
Keywords with Scores
Total Num Limit and Percentage of Named Entities
Extracted Keywords
Hypergraph Visualization
Chrono Graph
Topic Visualization Illustration

Fig. 2. Topic Visualization framework architecture

We summarize our primary contributions as follows:

- We introduce Chrono Clustering, a simple but powerful framework to figure out temporal changes of a topic.
- We use hypergraph to visualize the progression of trending topics, enabling people to intuitively grasp topic consistency and divergence over time.
- We employ an enhanced keyword extraction methodology to identify topic expression terms, allowing accurate description of evolving research concept trajectories.

II. Related Work

Various topic modeling techniques have emerged, each with unique capabilities for analyzing textual patterns. This section surveys key methods and their strengths and weaknesses.

Latent Dirichlet Allocation (LDA): LDA (Blei et al., 2003) extracts thematic structures by modeling topics as word mixtures. But LDA assumes topics are static and requires pre-setting the number of topics, limiting its handling of evolving text collections, while our approach could integrate temporal analysis to capture topic changes.

Latent Semantic Analysis (LSA): LSA (Landauer & Dumais, 1997) could identify word-text relationships through dimensionality reduction. However, as a method based on matrix factorization, LSA lacks interpretability and intuitiveness. Our method, instead, is able to extract topic expression terms and utilizes hypergraph to visualize dynamics of topics over time.

Non-negative Matrix Factorization (NMF): NMF (Lee & Seung, 1999) operates by breaking down data into two non-negative matrices, often referred to as the basis and coefficient matrices. This process allows for the extraction of features from the data, facilitating part-based representations where each data point is constructed from a combination of these features. NMF achieves this by iteratively updating the matrices to minimize the difference between the original data and the product of the basis and coefficient matrices. However, NMF treats all data points as if they are mutually independent, and ignores progression or changes over time. Thus, it cannot analyze temporal datasets which our method specially focuses on.

TextRank: TextRank (Mihalcea & Tarau, 2004) extracts keywords effectively with a graph-based ranking model, where nodes represent words, and edges reflect co-occurrences within text. Despite its efficacy in identifying salient words, TextRank may overlook overarching topics, as it primarily assesses the immediate context of words rather than their broader thematic relevance.

Continuous-Time Dynamic Topic Models (cDTM): cDTM (Blei & Lafferty, 2006) could capture topic evolution in text corpora over time. It applies variational Kalman filtering and smoothing to uncover thematic structures, assuming topic changes resemble Brownian motion. However, cDTM's computational load increases with larger datasets, limiting its use in broader applications due to scalability issue. Our lightweight dual clustering approach could readily handle large datasets.

Semantic Similarity Measures: Methods like cosine similarity or Jaccard index are very common but instrumental in improving the accuracy of topic clustering (Huang, 2008). These measures are straightforward, easy-to-use, and effective but may not capture deeper and contextual relationships of words in texts. Our method takes advantage of embeddings generated by the pre-trained transformer model on each piece of text to capture the essence of the text for the following topic clustering.

Hierarchical Dirichlet Process (HDP): HDP (Teh et al., 2006) uses a nonparametric Bayesian framework for modeling document collections, determining an optimal number of topics. HDP achieves this by iteratively assigning data points to topics, thus revealing a hierarchy of topics within the data. However, this process can be computationally intensive, and HDP cannot clearly distinguish between different topics when analyzing large datasets (Blei, Griffiths, & Jordan, 2010). In contrast, our technique combines scalability with temporal analysis to solve that issue.

Pachinko Allocation Model (PAM): PAM (Li & McCallum, 2006) analyzes topics by organizing them into a hierarchical structure and examining the connections between these levels. It achieves this by using a network of distributions that represent topics at different layers of the hierarchy, where each node in this network corresponds to a topic and edges represent the probabilistic relationships between them. Despite its modeling power, PAM's structural complexity requires sophisticated computational methods for accurate parameter estimation and can be challenging to interpret due to the model's depth and interconnectedness. On the contrary, our enhanced visualizations simplify topological analysis to make it easier to interpret.

In summary, while existing methods have advanced topic modeling, gaps remain in handling temporal data, identifying meaningful trends, computational efficiency, and semantic relationships. Our work presents a new technique to address these limitations and enable holistic dynamic topic analysis.

III. Methodology

We propose Chrono Clustering, a new temporal dual-clustering approach designed to uncover and visualize the dynamic evolution of topics over time. Chrono Clustering takes a text dataset across multiple timeframes as input. It can initiate analysis from any timeframe to follow-up developments of topics.

The method first leverages document embeddings (Mikolov et al.2013) and K-Means clustering (MacQueen, 1967) to identify key topics in the chosen timeframe. Within a given timeframe, Chrono Clustering identifies topics by grouping documents with similar meanings. This is done using document embeddings, which are numerical vector representations created by pre-trained transformer models (Vaswani, A. et al., 2006). These embeddings capture the essence of document contents, so when documents have similar embeddings, it suggests that they are related and collectively represent a latent topic. It then extracts Topic Cores and Radii as clustering anchors. A core innovation of Chrono Clustering is its use of these anchors to align texts in other time frames via K-Medoids (Kaufman & Rousseeuw,1987). This enables quantifying Topic Shifts from original Topic Cores. Another novel aspect is the recomputation of Topic Popularity within Topic Radii to capture trends of a topic over time.

Through integrated realignment, trend analysis, and customized visualization modules, Chrono Clustering provides a comprehensive framework to illuminate topic progression. In summary, Chrono Clustering introduces a new framework that combines tailored clustering, temporal coherence optimization, evolution analysis, and graphical modeling to holistically examine the dynamics of topics over time.

A. Document Embedding and Base Clustering

The first phase of Chrono Clustering is converting text documents into semantic embedding with techniques like Sentence Transformers (Reimers and Gurevych, 2019). This lets us effectively compare how similar the concepts are across documents.

We then use K-Means clustering on the vector embeddings from the base timeframe dataset, which could be any one of timeframe datasets based on the research deamand. This base clustering identifies the initial set of topics and their associated document groups. The cluster centers represent the 'Topic Core' for each topic. We also determine the 'Topic Radius', defined as the maximum distance between the Topic Core and any other points within its cluster, which captures the breadth of the topic. For each Topic Core, we further calculate the 'Topic Popularity' by aggregating the inverse distances of all points within the Topic Radius. This measures the density of the corresponding topic. Denser, more popular of the topic. In details, the popularity of a topic is calculated using the following formula:

$$Topic\ Popularity = \sum\nolimits_{Point\ within\ Topic\ Radius} \frac{1}{1 + Euclidean\ Distance(Topic\ Core, Point)}$$

The Topic Cores, Radii and Popularities extracted from base clustering give us a baseline for examining the evolution of these topics across different timeframes.

B. Temporal Realignment

Assume we start with the timeframe T_{base} to perform the base clustering and obtained N base topic clusters. The next step focuses on aligning these topics either to previous timeframes for tracing their origins and early developments or to following timeframes for extrapolating their further trajectories. In other words, the proposed temporal realignment is a longitudinal, retrospective-prospective approach to obtain a live-cycle view of those topics extracted from the timeframe T_{base}.

Technically, we apply one iteration of K-Medoids clustering to align the base topic clusters to documents in another timeframe T_x , with the topic cores from them as the medoids. More specifically, we group each document in T_x to the nearest medoids (topic cores from the base topic clusters) based on document embeddings. The newly formed clusters in T_x represent the reappearance of those base topics extracted from T_{base} . It is possible that no document is assigned to a medoid in T_x by one iteration of K-Medoids, which implies that there is no appearance of this topic in T_x. For each cluster obtained in T_x, we further exclude the medoid and then identify the real center of the cluster by computing the mean of the cluster. This center represents the focal point of this topic in T_x.

We then quantify the 'Topic Shift' as the Euclidean distance between this new center and the original Topic Core. This captures the degree of change the topic has undergone between T_x and T_{base}. Additionally, we recompute the Topic Popularity for the realigned cluster by considering only the points contained within the Topic Radius from the Base Clustering.

This reassessment situates Topic Popularity within the scope of the original topic, further enhancing temporal comparability. The Topic Shifts and changes in Topic Popularity provide the foundation for analyzing topic trends over time, as topics evolve through incremental deviations and sudden disruptions.

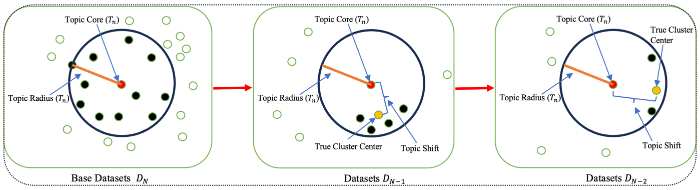

Fig. 3(A). Sketch Map: Backward Topic Progression in Temporal Realignment
In D_N, extract Topic Core, Topic Radius of the topic T_n by checking cluster points; Then apply them backward to previous datasets, the whole picture stands for the emergence and growth of a topic.

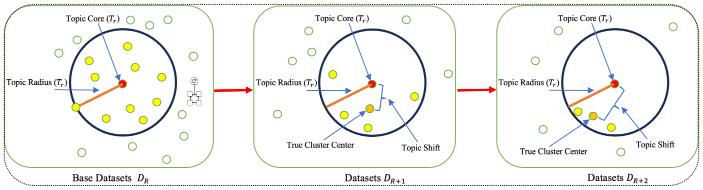

Fig. 3(B). Sketch Map: Forward Topic Progression in Temporal Realignment
In D_r, extract Topic Core, Topic Radius of the topic T_r by checking cluster points; Then apply them forward to later datasets, the whole picture stands for the diminishment of a topic.

We provide two illustrative examples of general topic development. As shown in Fig 3(A), it demonstrates the backward tracking of topic origin and development, revealing how current topics have emerged and grown gradually. Conversely, Figure 3(B) illustrates forward progression, highlighting the diminishment of topics over time.

C. Trend Analysis

With the Topic Shifts and recomputed Topic Popularities obtained through Temporal Realignment, Chrono Clustering analyzes these metrics to identify temporal trends.

We focus on topics that exhibit statistically significant changes in either Topic Shift or Popularity over time. These changes can indicate incremental progressions or radical disruptions in the topic's evolution. To identify significant changes, we could leverage techniques like:

- Thresholding: Defining threshold values for change magnitudes to classify as significant. For example, Topic Shifts above a certain Euclidean distance.
- Standard Deviation Analysis: Flagging changes exceeding a predefined number of standard deviations from the mean.
- Contextual Analysis: Assessing if changes are significant given the context of related topics and external developments.

The goal is to isolate topics displaying substantial shifts indicative of major trends, transitions, or anomalies. This provides the basis for deeper investigation into changes in trending directions.

Besides quantitative analysis, we also contextualize significant topic metric variations using qualitative techniques:

- Reviewing documents contributing to major shifts to explain swings.
- Consulting subject matter experts to validate technical mutations.
- Examining external events driving disruptions.

This blended approach combining computational methods and contextual review provides a rigorous framework for trend analysis on the evolving topic landscape.

D. Topic Visualization

For topics exhibiting significant evolutionary trends based on the Trend Analysis, Chrono Clustering visualizes the changes to illuminate topic progression over time. We generate visualizations called Chrono Graphs using a tailored keyword extraction technique and hypergraph representation.

First, we extract salient keywords from documents associated with a trending topic across different timeframes. To select semantically significant keywords, we leverage approaches like:

- TF-IDF Methodology with adjustments to overweight named entities and multi-word phrases
- Set the percentage of named entities in final keyword list to allow for semantic specialty
- Pruning common but semantically insignificant words

For keyword extraction, Adjusted TF-IDF Score is computed using the formula:

$$\begin{aligned} &\mathit{Adjusted\ Score}(\mathit{Named\ Entity}) \\ &\quad = \mathit{Original\ TF\_IDF\ Score} * \mathit{Length}(\mathit{Named\ Entity}) \\ &\quad * \mathit{Adjusted\ Factor}1 \end{aligned}$$

$$\begin{aligned} &\mathit{Adjusted\ Score}(\mathit{Phrase}) \\ &\quad = \mathit{Original\ TF\_IDF\ Score} * [1 + (\mathit{Length}(\mathit{Phrase}) \\ &\quad - 1) * \mathit{Adjusted\ Factor}2] \end{aligned}$$

This produces a timeline of keywords representing the topic evolution.

We then visualize these keywords in a hypergraph structure. Each timeframe forms a hyperedge, with keywords as vertices. The intersections between hyperedges denote evolutionary consistency, while disjoint keywords highlight temporal shifts. The resulting Chrono Graph provides an intuitive visualization of a topic's progression. Shared keywords represent its enduring core focus while disjoint terms reveal its changing facets over time. The visualization shows both incremental developments along stable trajectories as well as radical shifts marking new paradigms. This grants unique insights into the dynamics of topics uncovered by Chrono Clustering, shown in Fig 4.

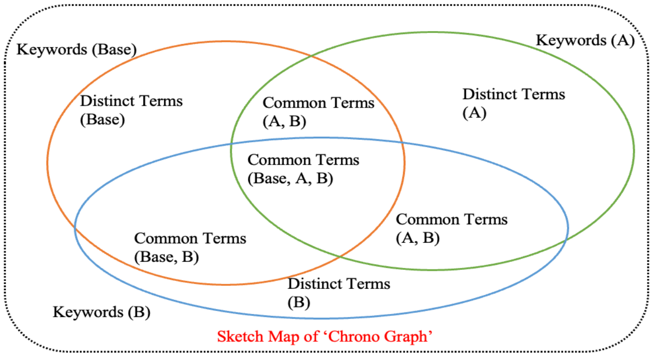

Fig. 4. Chrono Graph Sketch Map

E. Sample Applications of Chrono Clustering

Chrono Clustering could prove useful for fields where analyzing trends over time is crucial. For example, in financial investing, people generally need to analyze research reports and news across years, to dig out potential economic trends. Our method could capture temporal changes and extract trending topics. Similarly, in academia, identifying emerging research areas early on is often advantageous, as it enables greater output and achievements. Chrono Clustering could help us to grasp topic trends and predict potential topic emergence. Overall, the temporal clustering capabilities of our method make it suitable for applications requiring temporal trend analysis on text datasets.

IV. Experiments

The primary objective of this experimental section is to validate the effectiveness of Chrono Clustering. We aim to demonstrate its capability by analyzing and visualizing the evolution of topics over time, specifically in the field of AI research.

A. Datasets Description

The dataset for our experiments consists of 8,850 CVPR(Conference on Computer Vision and Pattern Recognition) abstracts and 6,101 ICML(International Conference on Machine Learning) abstracts from 2019 to 2023. Abstracts were chosen over full papers because they concisely summarize the core content in a more manageable format for analysis. As two premier AI conferences, CVPR and ICML serve as ideal sources for investigating our research questions. Data extraction was performed using BeautifulSoup to target paper titles, author names, and abstract texts.

B. Experimental Setup

- We first transformed all abstracts from ICML and CVPR datasets into embeddings with Sentence Transformer. Each embedding is a 768-dimensional vector and only corresponds to one abstract.
- With the embeddings prepared, we initiated 'Base Clustering' on the 2023 dataset using K-Means with K = 50, a moderate number, dividing the dataset into 50 clusters. And then identified 'Topic Cores' and establish 'Topic Radius'. Concurrently, we calculated the 'Topic Popularity' for each 'Topic Core'.
- 'Temporal Clustering Analysis' was then conducted on abstracts from 2019 to 2022. Using one-iteration K-Medoids, we aligned these abstracts to the Topic Cores defined in the base year(2023) dataset, tracking 'Topic Shift' and recalculating 'Topic Popularity' for each year.
- The ICML and CVPR datasets were processed separately to maintain the integrity of topic trends within their respective domains.
- Trending topics, discerned through significant changes in 'Topic Shift' and 'Topic Popularity,' were subsequently visualized via 'Chrono Graph', illustrating the temporal evolution of topics within each year.

C. Clustering Analysis

In this phase, we assessed significant changes in 'Topic Shift' and 'Topic Popularity' from clustering results to identify trending topics.

As shown in TABLE I and TABLE II, there are some topics with significant temporal changes in Topic Shift and Topic Popularity from our experiments. For topics with such changes, we identified them as potential trending topics. And we then needed to check their corresponding context datasets to figure out topic info.

By examining abstracts from 'Base Clustering' results, we identified specific burgeoning topics like 'Diffusion Models', 'Federated Learning', 'Contrastive Learning', and 'NeRF', which all align with currently surging AI research directions.

TABLE I. TEMPORAL DEVELOPMENT OF SAMPLE TOPICS IN CVPR

Topic (2023)	Topic Shift_2022	Topic Shift_2021	Topic Shift_2020	Topic Shift_2019	Topic Popularity_2023	Topic Popularity_2022	Topic Popularity_2021	Topic Popularity_2020	Topic Popularity_2019
NeRF	0.122881	0.26664	0.430557	0.508912	41.309243	25.983992	7.526164	1.790329	1.195469
Diffusion Models	0.316466	0.550796	0.685749	NaN	28.533831	5.437431	0	0.593208	0
Contrastive Learning	0.176677	0.269527	0.323347	0.454417	24.137339	12.868815	6.016083	3.601651	1.804946
Federated Learning	0.194781	0.386989	0.604594	NaN	16.004188	10.401417	2.985982	1.141489	0
Fast Time Adaptation	0.310762	0.73866	0.702865	NaN	11.462214	3.602704	0.575156	0.587246	0

In TABLE I, 'NaN' means that corresponding cluster is empty and Topic Shift cannot be figured out; '0' in Topic Popularity means no point within the corresponding Topic Radius.

TABLE II. TEMPORAL DEVELOPMENT OF SAMPLE TOPICS IN ICML

Topic (2023)	Topic Shift_2022	Topic Shift_2021	Topic Shift_2020	Topic Shift_2019	Topic Popularity_2023	Topic Popularity_2022	Topic Popularity_2021	Topic Popularity_2020	Topic Popularity_2019
Diffusion Models	0.442044	0.387999	0.640436	0.628758	21.586041	1.72883	2.303273	1.117622	1.127886
LLM	0.395634	0.338592	0.315765	0.383599	16.912067	2.318863	5.119087	5.105682	3.387843
AI in Molucular Structure	0.262256	0.277122	0.299828	0.473509	15.348213	6.075628	4.765512	5.339292	1.72186
Causality in AI	0.220503	0.330102	0.439262	0.447414	15.088233	8.180785	2.326602	2.31146	1.177772
Data Poisoning	0.40519	0.567742	0.532082	0.837625	9.986859	2.354797	0.617058	1.121324	0.544181
AI in Drug Discovery	0.284178	0.394805	0.580832	0.803279	10.032893	4.655189	2.922067	0.574552	0.554545
Quantum Computing	0.452749	0.393159	0.56875	0.576644	7.850368	1.764247	1.795158	0.569625	0.634259

D. Keyword Extraction and Visualization of Trending Topics

Then, we preprocessed abstracts from these trending topics, within the 'Topic Radius' using the nltk package (Bird, S., et al. 2009). This involves tokenization, lemmatization, and removal of stopwords, with a focus on named entities and general phrases. Capitalized words at sentence beginnings are excluded to prevent misidentification of named entities, for named entities were generally capitalized for the initial letter. Additionally, we have predefined a list of common words to exclude as shown below:
Predefined Word List = ['propose', 'method', 'task', 'approach', 'algorithm', 'system', 'technique', 'framework', 'performance', 'result', 'analysis', 'study', 'research', 'data', 'work', 'paper', 'findings', 'discussion', 'conclusion']

We then applied the TF-IDF method, to score each keyword. The scores were adjusted for named entities and longer phrases, with 'Adjusted Factor' of 1.5 and 0.1 respectively, recognizing that longer entity terms encode pivotal domain-specific concepts while moderately boosting meaningful multi-word expressions.

For the number of each final keyword list, we set 25 for each, the proportion of named entities as 60% and general phrases as 40%, both selected based on their adjusted TF-IDF score rankings. By doing this, we compiled final lists of keywords of different years within the same trending topic, and then visualized them with hypergraph methodology.

E. Results and Evaluation

In this part, we just randomly pick three graphs to validate the feasibility of Chrono Clustering.

In Fig 5, which showcases 'Diffusion Models', the absence of plots for 2019 and 2021 aligns with the real-world emergence of this topic around 2022. The 2020 plot reveals a reliance on traditional computer vision methods, while by 2022, 'Diffusion' related terms become prominent. The 2023 plot introduces new trends like 'Stable Diffusion' and 'Brownian Bridge', with overlapping areas indicating the topic's focus on image generation. This progression clearly encapsulates the actual trend of 'Diffusion Models'.

In Fig 6, as we could see from the core overlapping area of datasets, 'object' and 'ImageNet', which fits the truth that contrastive learning is generally used to train model to do object recognition, mainly from images. In 2019, saliency detection models were popular to train contrastive learning models. In 2020, research direction turned to ASRL and Video Object Grounding methods. In 2021, something breaking showed up, such as MoCo, a new framework based on Contrastive Learning. In 2022 and 2023, the research focus converted a new model CLIP developed by OpenAI, which was based on contrastive learning. Up to now, CLIP is still one of most popular AI models.

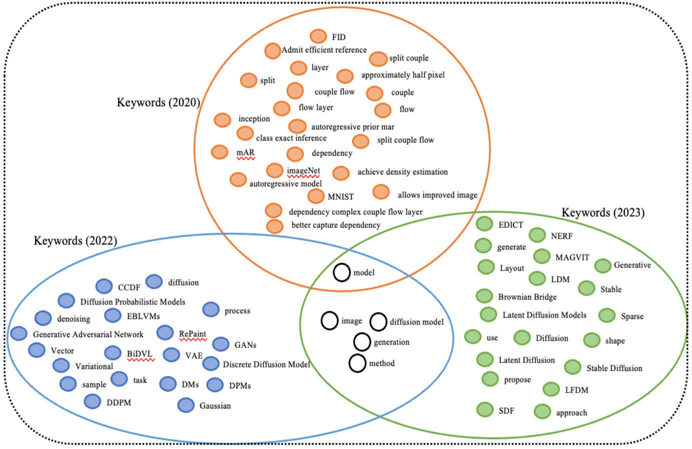

Fig. 5: It does not show keywords from 2019 and 2021, which means there are no related clusters in these two years. This graph shows the evolving trends of 'Diffusion Models'.

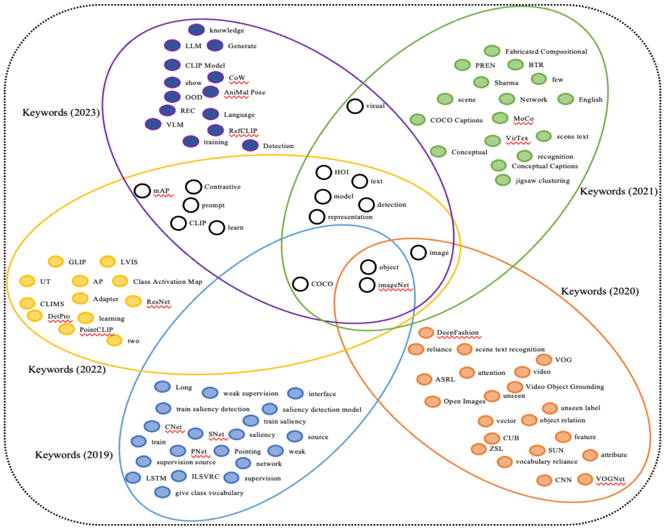

Fig. 6: This graph shows the evolving trends of 'Contrastive Learning'.

Similarly, Fig 7. The hypergraph outlines Federated Learning's advancement, starting in 2020 with an emphasis on 'Differential Privacy' for data protection. In 2021, the field solidifies with 'Episodic Learning' and 'Continuous Frequency Space' emerging, suggesting new angles in tackling the unique challenges of decentralized data. The year 2022 sees the introduction of novel frameworks, reflecting a broadening of the field's applications. By 2023, Federated Learning incorporates 'Model Distillation' and 'Sharpness Aware' methods, highlighting a progression towards refining models for enhanced performance and generalization in distributed environments. Also in the overlapping area, these terms are more obviously relevant, such as 'FL', 'global model', and 'Federated'.

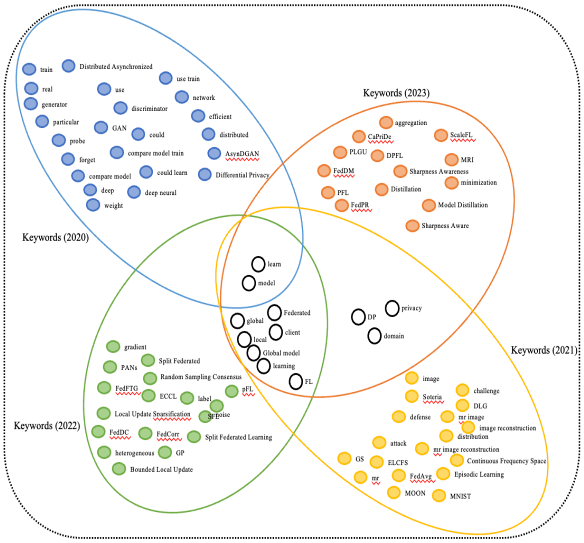

Fig. 7: This graph shows the evolving trends of 'Federated Learning.

The case studies validated the effectiveness of Chrono Clustering in discovering topic shifts. As demonstrated, these discovered topic shifts match real-world developments - emergence, growth, transitions, or dissipations of research topics in AI. Furthermore, Chrono Clustering produces intuitive visualizations showing changes in keywords describing the same topic over time. Our method allows users to better understand the inner dynamics of how topics are evolving.

V. Conclusion

In this paper, we introduced Chrono Clustering, an innovative clustering method for temporal text-based datasets. We defined a framework through K-Means and one-iteration K-Medoids to extract trending topics across different timeframes and designed an enhanced keyword extraction method with hypergraph to visualize. Throughout the experiments on two real-world datasets, we showed that our method demonstrates superior clustering performance, identifies trending topics, and generates intuitive 'Chrono Graph' that could be used for further analysis and decision-making.

While our method enhances temporal text analysis, we recognize the need for greater scalability and semantic depth in future developments. Enhancing language adaptability is also crucial for wider applicability.

Future Work. Future research directions for Chrono Clustering may explore its application across various domains, such as finance research, scientific literature analysis, and industry competitiveness tracking over time. There is also substantial potential to optimize the method for real-time usage, from early disease outbreak recognition to supply chain disruption anticipation and AI safety research prioritization, with the societal benefits spanning multiple critical frontiers.

References

[1] Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84.

[2] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022.

[3] Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse processes, 25(2-3), 259-284.

[4] Schofield, A., & Mimno, D. (2016). Comparing apples to apple: The effects of stemmers on topic models. Transactions of the Association for Computational Linguistics, 4, 287-300.

[5] Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review, 104(2), 211.

[6] Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755), 788-791.

[7] Xu, W., Liu, X., & Gong, Y. (2003). Document clustering based on non-negative matrix factorization. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (pp. 267-273).

[8] Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language processing (pp. 404-411).

[9] Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning (pp. 113-120).

[10] Huang, A. (2008). Similarity measures for text document clustering. In Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand (pp. 49-56).

[11] Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet processes. Journal of the American statistical association, 101(476), 1566-1581.

[12] Blei, D. M., Griffiths, T. L., & Jordan, M. I. (2010). The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. Journal of the ACM (JACM), 57(2), 1-30.

[13] Li, W., & McCallum, A. (2006, June). Pachinko allocation: DAG-structured mixture models of topic correlations. In Proceedings of the 23rd international conference on Machine learning (pp. 577-584).

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2013.

[15] MacQueen, J. (1967). "Some Methods for Classification and Analysis of Multivariate Observations," *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1, pp. 281-297.

[16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I. (2017). Attention Is All You Need. In Advances in Neural Information Processing Systems (pp. 5998-6008).

[17] Kaufman, L., & Rousseeuw, P. J. (1987). "Clustering by Means of Medoids," in *Statistical Data Analysis Based on the L1-Norm and Related Methods*, pp. 405-416.

[18] Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.

[19] Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media, Inc.