Event Hyper Graph

Sam Yu-Te Lee, Kwan-Liu Ma

Abstract—TBD

1 Introduction

In text analysis, providing a compact overview of a large collection of documents is a common yet challenging task. A large collection of documents, or a corpus, could contain multiple interesting characteristics, while the analyst may not already know what to expect. State-of-the-art techniques can be mainly divided into two categories: word clouds based and topic models based, each having its advantages and disadvantages. Word clouds are easy to understand and can be generated quickly, but they are not able to capture the relationships between words and support further analytical tasks [41]. Topic models can capture the relationships between words, but they are not easy to understand and require parameter tuning [8]. The ability to model complex relations lying within the text and interpretability are two important factors to consider when providing a compact overview of a corpus, and they are often in conflict with each other.

In addition, a common assumption shared among text analysis methods (not limited to word clouds and topic models) is that each document is a bag of words, and words appearing near each other indicate meaningful relations. There are two main problems with this assumption. First, the fact that meaningful relations between words are informed solely by their co-occurring frequency indicates that the relation is unknown. Co-occurring frequency merely serves as a signal for the possibility that a relation exists. Many visual analytics systems did not convey this weak signaling. Instead, many systems assume the extracted relations already have meanings. They then visualize the extracted relations for the user to interpret the exact meanings they convey. This misinformed visualization thus leads to misinterpretation [27]. Second problem is that the meaning hidden under co-occurrence frequency either does not exist at all or is not to the user's interest in most cases [16]. For example, Chun et al. [10] found in a biomedical text corpus that only 30% of protein pairs co-occurring in the same sentences have an actual interconnection. When looking for a specific relation, co-occurrence cannot provide much help, and thus any methods that build upon co-occurrence would inevitably fail.

Event hypergraph can capture complex relations while being easy to understand. It utilizes the power of event extraction models to capture the complex relations in the text and uses hypergraphs to visualize the result and support further analytical tasks. The extracted relations are expressed as *events* rather than co-occurrences, and the semantic meanings of the relations are explicitly represented by the event through its type, trigger and arguments, which are all human-readable text mentions. By organizing events into a hypergraph, we provide a compact overview of the corpus in a graph where events and words are interconnected. The graph representation allows users to interact with the overview through common interactions such as filtering, zooming and highlighting. It also opens the door to incorporating hypergraph analysis methods into text analysis, such as community detection and motif finding. A study by Antelmi et al. [2] found that in certain scenarios where relations could exist between more than two nodes, conducting analysis on hyper networks can provide more accurate results than on regular networks. In the case of event hypergraphs, the nodes (entities) and hyperedges (events) are all textual mentions extracted from the documents. This representation preserves the syntactic and semantic information of the text, enabling the user to easily verify the validity of the extracted relations. Conducting hypergraph analysis on such a network has not been previously done by the literature. The next step of this project is to find appropriate case studies to employ hypergraph analysis methods and develop appropriate interaction support.

Our contributions are:

- a novel method that combines event extraction models and hypergraphs to provide a compact, interactive overview of a large collection of documents.
- a system that combines our document overview with hypergraph analysis methods for one to explore and analyze a large collection of documents.
- two case studies that demonstrate the effectiveness, interpretability, and trustworthiness of Event Hypergraph.

2 Related Works

2.1 Corpus level Overview

2.1.1 Word Clouds

Word Clouds are intuitive and effective means for users even with the lowest visual literacy to understand the main topics in a collection of documents instantly. The most basic word clouds display the most frequent words in the collection in a visually appealing way, with the size of the words representing their frequency. More advanced word clouds try to enhance it by incorporating more information, such as temporal attributes [23, 24], relations between words [11, 21] and semantics [12, 35, 42, 43]. However, regardless of the enhancements, common limitations of the word clouds exist, and some are at the root of the statistical assumptions of word clouds.

First, word clouds assume the independence between words, and all statistical calculations such as word frequency and co-occurrence are based on this assumption. This assumption leads to the inability to preserve complex relations between words in the text and eventually limits its visual representation. Since all words are treated independently, the position of each word is usually randomized in the visualization of a word cloud. This randomization is not a design choice, but the implication of not being able to capture anything meaningful to be encoded as positions. Semantic-preserving word clouds try to address this issue by grouping semantically similar words, and weakly encoding the position of each word, but the relative position of words in each group is still randomized.

The second limitation of word clouds is the lack of support for further analytical tasks [41]. This is due to some common issues of their visual encoding. For example, the size encoding of each word is not accurate due to different word lengths, making it difficult to compare words. Despite its popularity, word clouds used in visual analytics systems are often used as an exploratory starting point for the user to select words, providing a static and non-interactive overview, while the rest of the system designs that support further analytical tasks on the selected words are not directly informed by the information provided by the word clouds. This decoupling of the word clouds and the rest of the system means that word clouds are better used

• *Sam Yu-Te Lee and Kwan-Liu Ma are with University of California, Davis. E-mail: {ytlee | klma}@ucdavis.edu.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

as a complementary visualization component, rather than the main component for exploratory analysis.

2.1.2 Topic Models

Topic modeling is a popular statistical tool for extracting latent variables (topics) from large collections of documents [40]. By making different assumptions and using different statistical models, an extensive amount of topic models are proposed, such as Latent Dirichlet Allocation (LDA) [4], Correlated Topic Model (CTM) [3], Pachinko Allocation Model (PAM) [28], Non-negative Matrix Factorization (NMF) [25], and their countless variations. Although having the ability to capture various complex relations in the text, the extensive amount of possible choices and their complex assumptions and modeling processes make it difficult for even the experts to choose the right model for their task. Moreover, each model might have hyperparameters that need to be tuned, which is often time-consuming and requires expert knowledge.

Addressing the complexity of topic models is not a trivial task. The complexity of the topic models even spurs the idea of using visual analytic systems to help users understand and refine the topic models [6, 7, 13, 17, 22, 26]. Despite extensive studies, Lee et al. found that non-expert topic model users constantly misinterpret the results of topic models [27]. When presented with common topic modeling results, users sometimes overlooked important words, read too much into words, or assumed adjacent words went together. Another work by Chuang et al. [9] proposes to evaluate topic modeling systems by *interpretation* and *trust*, Based on this evaluation, they conducted a literature review and finds that most tools lack proper consideration of how model abstractions align with analysis tasks, thus lacking interpretability and trustworthiness

2.1.3 Text Patterns

TBD

2.1.4 Entity based methods

Besides the majority of works that use word clouds and topic models to provide an overview at the corpus level, some other works also use various methods to summarize a corpus.

ConceptVector [34] proposes a user-steerable word-to-concept similarities model where the user can define and refine concepts, which are then used to organize document corpus and conduct analysis. Their similarity computation is based on word embeddings but enhanced to compute similarity (relevance) between a concept and a document. As an earlier approach, Jigsaw [36] extracts entities mentioned in each document and connects entities if they appear in the same document. It provides a list view as well as a graph view, where each document is a node and the mentioned entities are connected to the document. This graph visualization approach is similar to what we use in this proposal, but instead of connecting entities by documents, we connect entities by events they participate in. FacteAtlas [5] also build upon the idea of organizing documents by entities. They first extract entities mentioned in a document along with their classes (facets) and then organize the corpus by facets. The result is represented by a multi-layer graph and visualized in a novel way. These approaches all assume that words co-occurring with each other have semantic meaningful relations, which is not always true as will be mentioned in subsection 2.2.

2.1.5 Embedding based methods

Recent approaches, such as VITALITY [30], uses document embeddings and dimensionality reduction methods to visualize the corpus in a scatter plot, where each circle represents a document. Despite being intuitive, these embedding-based approaches lack interpretability, and it is not straightforward for the user to understand why similar documents are grouped together.

2.2 Event Extraction

To capture an accurate relation between words, researchers from Natural Language Processing (NLP) community have developed *event extraction* (EE) methods [16]. In this context, the most basic definition of an *event* is defined as a structure consisting of a *trigger* and zero or more *arguments*, where the *trigger* is the textual mention of verbs that clearly expresses the occurrence of an event, and the *arguments* are the textual mentions of the participants of the event, usually a named entity. Additionally, each argument typically has an associated *argument role*. For example, taking the input: *'The man returned to Los Angeles from Mexico following his capture Tuesday by bounty hunters.'*, an event extraction model should extract *'return'* as the trigger and *'man', 'Los Angeles', 'Mexico', 'Tuesday' and 'bounty hunters'* as the arguments. The arguments are thus semantically differentiated by their argument roles. In slightly more advanced definitions, events might also have *event property*, and events could be nested, where an argument of an event is another event. Event trigger and its arguments are not necessarily in the same sentence, thus the extracted relations do not rely on the aforementioned bag-of-words assumption. In addition, relations between words are represented by the trigger, enabling the possibility of capturing diverse but interpretable relations between words.

Event extraction (EE) is a well-studied problem in the NLP community, and various methods have been proposed. To begin with, EE models can be divided into *Close Domain* (CDEE) and *Open Domain* (ODEE) models. CDEE models use a predefined event schema, which defines targeting event types and corresponding argument types, and the goal is to fill the expected 'slots'. In the aforementioned example, two events can be extracted: The man *returning* and the man being *captured*. Using a predefined event schema enables better accuracy on the targeted events by ignoring other events. This trade-off is constantly being made in CDEE models to account for different application scenarios. As a result, CDEE solutions are usually limited to a specific domain. Their labeled dataset scale is also limited because event schemas are not easily transferrable, requiring different labels for different schemas. On the other hand, ODEE models do not assume a predefined event schema, and the goal is to extract all possible events from the text. Though promising, ODEE models are rare, and model evaluation can only be done manually.

Early approaches to EE adopted human-crafted templates to match the events. These rule-based pattern-matching approaches are promised to be accurate if the templates are well-designed, it is expensive to build and maintain and hardly generalizable. Recent approaches use supervised deep learning models to extract events. Despite various variations, most deep learning models use popular datasets like Cancer Genetics 2023 (CG), GENIA 2011/2013 (GE11, GE13), and Infectious Disease (ID), and fine-tune Bert-based or similar language models to extract events. Deep learning models achieve state-of-the-art (SOTA) performance, but they are usually CDEE models, and thus the model can only apply to the domain of the training data. To the best of our knowledge, most datasets are in either the biomedical or news domain, limiting the event extraction models to these domains. In the biomedical domain, the SOTA model is DeepEventMine [38], and in the news domain, it is Text2Event [29]. In our work, we test these two models in two different case studies to prove the efficacy of using event extraction models.

2.3 Hyper Graph Visualization

A hypergraph is a generalization of a graph, where an edge can connect more than two nodes [15]. The result of event extraction can be transformed into a hypergraph, where an event is a hyperedge, and the arguments are the nodes. Traditionally, hypergraphs are visualized as Venn diagrams or Euler diagrams, but these visualizations are limited to small hypergraphs. More advanced visualizations can be categorized into node-link based, matrix based, and timeline based approaches.

In node-link based approaches, hyperedges are represented as an extra node, and the hypergraph is represented as a bipartite graph. The resulting graph is similar to a heterogeneous graph and can be visualized using normal graph layout algorithms. Node-link based approach is the most common approach and is used in many hypergraph visualization systems [18–20, 32, 33]. Kapec [19] and Paquette and Tokuyasu. [33] first use the extra node representation, and Ouvard et al. [32] refine the visualization by minimizing the size of hyper-edge nodes. Node-link based approaches are the most intuitive and have the best support

for interaction, but the visual scalability is limited as all graph layout algorithms.

The timeline based approaches target dynamic hypergraphs, and usually treat the hypergraph visualization problem as a set membership visualization problem [1,31,39]. At each time step, each hyperedge can be seen as a subset of the set of nodes, and the goal is to visualize the membership of each node at different time steps. For example, Agarwal et al.use a Sankey diagram to visualize dynamic hypergraphs, where each subset (hyperedge) is represented by a vertical bar in the Sankey diagram and the nodes are represented by flows [1].

Matrix-based approaches [14, 37] put nodes and hyperedges on the rows and columns, and use the cell to encode where the node belongs to a hyperedge. The benefit of using a matrix-based visualization is scalability and better support for dynamic networks. At this point, there are still many unsolved problems to use a matrix-based visualization for hypergraphs. First, the row and column ordering of the matrix visualization is important to reveal interesting patterns. Also, matrix visualization assumes that hyperedges do not have any relation; this prevents visualizing hypergraphs with nested events. Finally, it is not trivial to visualize the temporal aspect of the hypergraph. Further research is needed to solve these problems.

Acknowledgments

The authors wish to thank A, B, and C. This work was supported in part by a grant from XYZ (# 12345-67890).

References

[1] S. Agarwal and F. Beck. Set streams: Visual exploration of dynamic overlapping sets. In *Computer Graphics Forum*, vol. 39, pp. 383–391. Wiley Online Library, 2020. 3

[2] A. Antelmi, G. Cordasco, B. Kamiński, P. Prałat, V. Scarano, C. Spagnuolo, and P. Szufel. Analyzing, exploring, and visualizing complex networks via hypergraphs using simplehypergraphs. jl. *arXiv preprint arXiv:2002.04654*, 2020. 1

[3] D. Blei and J. Lafferty. Correlated topic models. *Advances in neural information processing systems*, 18:147, 2006. 2

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003. 2

[5] N. Cao, J. Sun, Y.-R. Lin, D. Gotz, S. Liu, and H. Qu. Facetatlas: Multifaceted visualization for rich text corpora. *IEEE transactions on visualization and computer graphics*, 16(6):1172–1181, 2010. 2

[6] A. Chaney and D. Blei. Visualizing topic models. In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 6, pp. 419–422, 2012. 2

[7] J. Choo, C. Lee, C. K. Reddy, and H. Park. Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE transactions on visualization and computer graphics*, 19(12):1992–2001, 2013. 2

[8] J. Chuang, S. Gupta, C. Manning, and J. Heer. Topic model diagnostics: Assessing domain relevance via topical alignment. In *International conference on machine learning*, pp. 612–620. PMLR, 2013. 1

[9] J. Chuang, D. Ramage, C. Manning, and J. Heer. Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 443–452, 2012. 2

[10] H.-W. Chun, Y. Tsuruoka, J.-D. Kim, R. Shiba, N. Nagata, T. Hishiki, and J. Tsujii. Extraction of gene-disease relations from medline using domain dictionaries and machine learning. In *Biocomputing 2006*, pp. 4–15. World Scientific, 2006. 1

[11] C. Collins, F. B. Viegas, and M. Wattenberg. Parallel tag clouds to explore and analyze faceted text corpora. In *2009 IEEE Symposium on Visual Analytics Science and Technology*, pp. 91–98, 2009. doi: 10.1109/VAST.2009.5333443 1

[12] W. Cui, Y. Wu, S. Liu, F. Wei, M. X. Zhou, and H. Qu. Context preserving dynamic word cloud visualization. In *2010 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 121–128. IEEE, 2010. 1

[13] M. El-Assady, R. Sevastjanova, F. Sperrle, D. Keim, and C. Collins. Progressive learning of topic modeling parameters: A visual analytics framework. *IEEE transactions on visualization and computer graphics*, 24(1):382–391, 2017. 2

[14] M. T. Fischer, D. Arya, D. Streeb, D. Seebacher, D. A. Keim, and M. Worring. Visual analytics for temporal hypergraph model exploration. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):550–560, 2020. 3

[15] M. T. Fischer, A. Frings, D. A. Keim, and D. Seebacher. Towards a survey on static and dynamic hypergraph visualizations. In *2021 IEEE visualization conference (VIS)*, pp. 81–85. IEEE, 2021. 2

[16] G. Frisoni, G. Moro, and A. Carbonaro. A survey on event extraction for natural language understanding: Riding the biomedical literature wave. *IEEE Access*, 9:160721–160757, 2021. doi: 10.1109/ACCESS.2021.3130956 1, 2

[17] B. Gretarsson, J. O'donovan, S. Bostandjiev, T. Höllerer, A. Asuncion, D. Newman, and P. Smyth. Topicnets: Visual analysis of large text corpora with topic modeling. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(2):1–26, 2012. 2

[18] B. Jacobsen, M. Wallinger, S. Kobourov, and M. Nöllenburg. Metrosets: Visualizing sets as metro maps. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1257–1267, 2020. 2

[19] P. Kapec. Visualizing software artifacts using hypergraphs. In *Proceedings of the 26th Spring Conference on Computer Graphics*, pp. 27–32, 2010. 2

[20] A. Kerren and I. Jusufi. A novel radial visualization approach for undirected hypergraphs. In *EuroVis (Short Papers)*, 2013. 2

[21] K. Kim, S. Ko, N. Elmqvist, and D. S. Ebert. Wordbridge: Using composite tag clouds in node-link diagrams for visualizing content and relations in text corpora. In *2011 44th Hawaii International Conference on System Sciences*, pp. 1–8, 2011. doi: 10.1109/HICSS.2011.499 1

[22] M. Kim, K. Kang, D. Park, J. Choo, and N. Elmqvist. Topiclens: Efficient multi-level visual topic exploration of large-scale document collections. *IEEE transactions on visualization and computer graphics*, 23(1):151–160, 2016. 2

[23] J. Knittel, S. Koch, and T. Ertl. Pyramidtags: Context-, time- and word order-aware tag maps to explore large document collections. *IEEE Transactions on Visualization and Computer Graphics*, 27(12):4455–4468, 2021. doi: 10.1109/TVCG.2020.3010095 1

[24] B. Lee, N. H. Riche, A. K. Karlson, and S. Carpendale. Sparkclouds: Visualizing trends in tag clouds. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1182–1189, 2010. doi: 10.1109/TVCG.2010.194 1

[25] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999. 2

[26] H. Lee, J. Kihm, J. Choo, J. Stasko, and H. Park. ivisclustering: An interactive visual document clustering via topic modeling. In *Computer graphics forum*, vol. 31, pp. 1155–1164. Wiley Online Library, 2012. 2

[27] T. Y. Lee, A. Smith, K. Seppi, N. Elmqvist, J. Boyd-Graber, and L. Findlater. The human touch: How non-expert users perceive, interpret, and fix topic models. *International Journal of Human-Computer Studies*, 105:28–42, 2017. 1, 2

[28] W. Li and A. McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*, pp. 577–584, 2006. 2

[29] Y. Lu, H. Lin, J. Xu, X. Han, J. Tang, A. Li, L. Sun, M. Liao, and S. Chen. Text2event: Controllable sequence-to-structure generation for end-to-end event extraction. *arXiv preprint arXiv:2106.09232*, 2021. 2

[30] A. Narechania, A. Karduni, R. Wesslen, and E. Wall. Vitality: Promoting serendipitous discovery of academic literature with transformers & visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):486–496, 2022. doi: 10.1109/TVCG.2021.3114820 2

[31] P. H. Nguyen, K. Xu, R. Walker, and B. W. Wong. Timesets: Timeline visualization with set relations. *Information Visualization*, 15(3):253–269, 2016. 3

[32] X. Ouvrard, J.-M. L. Goff, and S. Marchand-Maillet. Networks of collaborations: Hypergraph modeling and visualisation. *arXiv preprint arXiv:1707.00115*, 2017. 2

[33] J. Paquette and T. Tokuyasu. Hypergraph visualization and enrichment statistics: how the egan paradigm facilitates organic discovery from big data. In *Human Vision and Electronic Imaging XVI*, vol. 7865, pp. 126–143. SPIE, 2011. 2

[34] D. Park, S. Kim, J. Lee, J. Choo, N. Diakopoulos, and N. Elmqvist. Conceptvector: Text visual analytics via interactive lexicon building using word embedding. *IEEE transactions on visualization and computer graphics*, 24(1):361–370, 2017. 2

[35] F. V. Paulovich, F. M. Toledo, G. P. Telles, R. Minghim, and L. G. Nonato. Semantic wordification of document collections. In *Computer Graphics*

Forum, vol. 31, pp. 1145–1153. Wiley Online Library, 2012. 1

[36] J. Stasko, C. Gorg, Z. Liu, and K. Singhal. Jigsaw: supporting investigative analysis through interactive visualization. In *2007 IEEE Symposium on Visual Analytics Science and Technology*, pp. 131–138. IEEE, 2007. 2

[37] D. Streeb, D. Arya, D. A. Keim, and M. Worring. Visual analytics framework for the assessment of temporal hypergraph prediction models. In *Set Visual Analytics Workshop at IEEE VIS 2019*, 2019. 3

[38] H.-L. Trieu, T. T. Tran, K. N. Duong, A. Nguyen, M. Miwa, and S. Ananiadou. Deepeventmine: end-to-end neural nested event extraction from biomedical texts. *Bioinformatics*, 36(19):4910–4917, 2020. 2

[39] P. Valdivia, P. Buono, C. Plaisant, N. Dufournaud, and J.-D. Fekete. Analyzing dynamic hypergraphs with parallel aggregated ordered hypergraph visualization. *IEEE transactions on visualization and computer graphics*, 27(1):1–13, 2019. 3

[40] I. Vayansky and S. A. Kumar. A review of topic modeling methods. *Information Systems*, 94:101582, 2020. 2

[41] F. B. Viégas and M. Wattenberg. Timelines tag clouds and the case for vernacular visualization. *interactions*, 15(4):49–52, 2008. 1

[42] J. Wang, J. Zhao, S. Guo, C. North, and N. Ramakrishnan. Recloud: semantics-based word cloud visualization of user reviews. In *Graphics Interface 2014*, pp. 151–158. AK Peters/CRC Press, 2020. 1

[43] Y. Wu, T. Provan, F. Wei, S. Liu, and K.-L. Ma. Semantic-preserving word clouds by seam carving. In *Computer Graphics Forum*, vol. 30, pp. 741–750. Wiley Online Library, 2011. 1