Event Hyper Graph
Proposal

Sam Yu-Te Lee

May 24, 2023

1 Motivation

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 own 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 for further analytical tasks [40]. Topic models can capture the relationships between words, but they are not easy to understand and require parameter tuning [7]. The ability to model complex relations lie 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. Word clouds and topic models lie on the two extremes: word clouds are easy to understand but too simple to capture complex relations, and topic models are able to capture complex relations but not easy to understand.

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 to 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 system did not clearly convey this weak signaling. Instead, many systems assumes the extracted relations already have meanings. They then visualize the extracted relations for users to interpret the exact meanings they convey. This misinformed visualization thus leads to misinterpretation [26]. 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 [15]. For example, Chun et al. [9] 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 can not provide much help, and thus any methods that build upon co-occurrence will inevitably fail.

The proposed method, event hyper graph, is able to 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 hyper graphs to visualize the result and support further analytical tasks. The extracted relations are expressed as *events* rather than co-occurrences, and the semantic meaning 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 hyper graph, we provide a compact overview of the corpus in a graph where events and words are interconnected. The graph form representation allows users to interact with the overview through common interactions such as filtering, zooming and highlighting. It also opens the door to incorporate hyper network analysis methods into text analysis, such as community detection and motif finding.

Our contributions are:

- We propose a novel method that combines event extraction models and hyper graphs to provide a compact, interactive overview of a large collection of documents that is both interpretable and trustworthy.

- We develop a system that combines the proposed overview with hyper network analysis methods for user to explore a corpus.
- We present two case studies to demonstrate the effectiveness of our method.

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 [22, 23], relations between words [10, 20] and semantics [42, 34, 41, 11]. However, regardless of the enhancements, common limitations of the word clouds exist, and some are in 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 together, and weakly encodes 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 [40]. This is due to some common issues of the visual encoding. For example, the size encoding of each word is not accurate due to different word length, 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 user to select words, providing a static and non-interactive overview, while the rest of the system design that supports 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 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 [39]. By making different assumptions and using different statistical models, an extensive amount of topic models are proposed, such as Latent DirichLet Allocation (LDA) [3], Correlated Topic Model (CTM) [2], Pachinko Allocation Model (PAM) [27], Non-negative Matrix Factorization (NMF) [24], and their countless variations. Although having the ability to capture various complex relations in the text, the extensive amount of possible choices and their each individual 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 hyper parameters 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 [12, 6, 25, 21, 5, 16]. Despite extensive studies, Lee et al. found that non-expert topic model users constantly misinterpret the results of topic models [26]. 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. [8] 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

2.2 Text Patterns

2.2.1 Others

Besides the majority of works that use word clouds and topic models to provide overview at the corpus level, some other works also use various methods to summarize a corpus. ConceptVector [33] proposes a user-steerable word-to-concept similarities model where 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. Despite being intuitive, these embedding-based approaches lacks interpretability, and it is not straightforward for the user to understand why similar documents are grouped together. As an earlier approach, Jigsaw [35] 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, which is a finer granularity. FacteAtlas [4] 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 assumes that words co-occurring with each other have semantic meaningful relations, which is not always true as will be mentioned in subsection 2.3.

Recent approaches, such as VITALITY [29], 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 lacks interpretability, and it is not straightforward for the user to understand why similar documents are grouped together.

2.3 Event Extraction

To capture an accurate relation between words, researchers from Natural Language Processing (NLP) community have developed *event extraction* (EE) methods [15]. In this context, the most basic definition of an *event* is defined as a structure consists 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. In slightly more advanced definitions, events might also include *argument roles* and *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 does 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 while interpretable relations between words.

Event extraction (EE) is a well-studied problem in 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 uses predefined event schemas, where it defines targeting event types and corresponding argument types, and the goal is to fill the expected ‘slots’. CDEE solutions are usually limited to a specific domain and require labeled training data. On the other hand, ODEE models does not assume predefined event schemas, and the goal is to extract all possible events from the text. Though promising, ODEE models are rare, and its evaluation can only be done manually.

Early approaches on 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), Infectious Disease (ID), and fine-tunes 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 biomedical or news domain, limiting the event extraction models to these domains. In biomedical domain, the SOTA model is DeepEventMine [37], and in news domain it is Text2Event [28]. We propose to test these two models in two different case studies to prove the efficacy of using event extraction models.

2.4 Hyper Graph Visualization

Hyper graph is a generalization of graph, where an edge can connect more than two nodes [14]. The result of event extraction can be transformed into a hyper graph, where an event is a hyper edge, and the arguments are the nodes. Traditionally, hyper graphs are visualized as Venn diagrams or Euler diagrams, but these visualizations are limited to small hyper graphs. More advanced visualizations can be categorized into node-link based, matrix-based, and timeline-based approaches.

In node-link based approaches, hyper edges are represented as an extra node, and the hyper graph is represented as a bipartite graph. The resulting graph is similar to a heterogeneous graph, and can be visualized using normal graph layout algorithms. This is the most common approach, and is used in many hyper graph visualization systems [18, 32, 19, 31, 17]. Kapec [18] and Paquette et al. [32] propose the earlier extra node approaches, and Ouvard et al. [31] further 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.

Timeline-based approaches targets dynamic hyper graphs, and usually treat the hyper graph visualization problem as a set membership visualization problem [1, 30, 38]. At each time step, each hyper edge 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.used a sankey diagram to visualize dynamic hyper graphs, where each subset (hyper edge) is represented by a bar in the sankey diagram, and the nodes are represented by flows [1, **?**].

Matrix-based approaches [36, 13] put nodes and hyper edges on the rows and columns, and use the cell to encode where the node belongs to a hyper edge. The benefit of using a matrix-based visualization is the scalability. However, similar to normal graphs, the row and column ordering of the matrix visualization is important to reveal interesting patterns. Also, matrix visualization assumes that hyper edges do not have any relation, this prevents to visualize hyper graphs with nested events. As a result, we will focus on node-link based approaches in this project.

TODO: Investigate tasks on Hyper Graphs

References

[1] Shivam Agarwal and Fabian Beck. Set streams: Visual exploration of dynamic overlapping sets. In *Computer Graphics Forum*, volume 39, pages 383–391. Wiley Online Library, 2020.

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

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

[4] Nan Cao, Jimeng Sun, Yu-Ru Lin, David Gotz, Shixia Liu, and Huamin Qu. Facetatlas: Multifaceted visualization for rich text corpora. *IEEE transactions on visualization and computer graphics*, 16(6):1172–1181, 2010.

[5] Allison Chaney and David Blei. Visualizing topic models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 6, pages 419–422, 2012.

[6] Jaegul Choo, Changhyun Lee, Chandan K Reddy, and Haesun Park. Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE transactions on visualization and computer graphics*, 19(12):1992–2001, 2013.

[7] Jason Chuang, Sonal Gupta, Christopher Manning, and Jeffrey Heer. Topic model diag-

nostics: Assessing domain relevance via topical alignment. In *International conference on machine learning*, pages 612–620. PMLR, 2013.

[8] Jason Chuang, Daniel Ramage, Christopher Manning, and Jeffrey Heer. Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 443–452, 2012.

[9] Hong-Woo Chun, Yoshimasa Tsuruoka, Jin-Dong Kim, Rie Shiba, Naoki Nagata, Teruyoshi Hishiki, and Jun'ichi Tsujii. Extraction of gene-disease relations from medline using domain dictionaries and machine learning. In *Biocomputing 2006*, pages 4–15. World Scientific, 2006.

[10] Christopher Collins, Fernanda B. Viegas, and Martin Wattenberg. Parallel tag clouds to explore and analyze faceted text corpora. In *2009 IEEE Symposium on Visual Analytics Science and Technology*, pages 91–98, 2009.

[11] Weiwei Cui, Yingcai Wu, Shixia Liu, Furu Wei, Michelle X Zhou, and Huamin Qu. Context preserving dynamic word cloud visualization. In *2010 IEEE Pacific Visualization Symposium (PacificVis)*, pages 121–128. IEEE, 2010.

[12] Mennatallah El-Assady, Rita Sevastjanova, Fabian Sperrle, Daniel Keim, and Christopher Collins. Progressive learning of topic modeling parameters: A visual analytics framework. *IEEE transactions on visualization and computer graphics*, 24(1):382–391, 2017.

[13] Maximilian T Fischer, Devanshu Arya, Dirk Streeb, Daniel Seebacher, Daniel A Keim, and Marcel Worring. Visual analytics for temporal hypergraph model exploration. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):550–560, 2020.

[14] Maximilian T Fischer, Alexander Frings, Daniel A Keim, and Daniel Seebacher. Towards a survey on static and dynamic hypergraph visualizations. In *2021 IEEE visualization conference (VIS)*, pages 81–85. IEEE, 2021.

[15] Giacomo Frisoni, Gianluca Moro, and Antonella Carbonaro. A survey on event extraction for natural language understanding: Riding the biomedical literature wave. *IEEE Access*, 9:160721–160757, 2021.

[16] Brynjar Gretarsson, John O'donovan, Svetlin Bostandjiev, Tobias Höllerer, Arthur Asuncion, David Newman, and Padhraic Smyth. Topicnets: Visual analysis of large text corpora with topic modeling. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(2):1–26, 2012.

[17] Ben Jacobsen, Markus Wallinger, Stephen Kobourov, and Martin Nöllenburg. Metrosets: Visualizing sets as metro maps. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1257–1267, 2020.

[18] Peter Kapec. Visualizing software artifacts using hypergraphs. In *Proceedings of the 26th Spring Conference on Computer Graphics*, pages 27–32, 2010.

[19] Andreas Kerren and Ilir Jusufi. A novel radial visualization approach for undirected hypergraphs. In *EuroVis (Short Papers)*, 2013.

[20] KyungTae Kim, Sungahn Ko, Niklas Elmqvist, and David 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*, pages 1–8, 2011.

[21] Minjeong Kim, Kyeongpil Kang, Deokgun Park, Jaegul Choo, and Niklas 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.

[22] Johannes Knittel, Steffen Koch, and Thomas 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.

[23] Bongshin Lee, Nathalie Henry Riche, Amy K. Karlson, and Sheelash Carpendale. Sparkclouds: Visualizing trends in tag clouds.

IEEE Transactions on Visualization and Computer Graphics, 16(6):1182–1189, 2010.

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

[25] Hanseung Lee, Jaeyeon Kihm, Jaegul Choo, John Stasko, and Haesun Park. ivisclustering: An interactive visual document clustering via topic modeling. In *Computer graphics forum*, volume 31, pages 1155–1164. Wiley Online Library, 2012.

[26] Tak Yeon Lee, Alison Smith, Kevin Seppi, Niklas Elmqvist, Jordan Boyd-Graber, and Leah Findlater. The human touch: How non-expert users perceive, interpret, and fix topic models. *International Journal of Human-Computer Studies*, 105:28–42, 2017.

[27] Wei Li and Andrew McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*, pages 577–584, 2006.

[28] Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. Text2event: Controllable sequence-to-structure generation for end-to-end event extraction. *arXiv preprint arXiv:2106.09232*, 2021.

[29] Arpit Narechania, Alireza Karduni, Ryan Wesslen, and Emily Wall. Vitality: Promoting serendipitous discovery of academic literature with transformers & visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):486–496, 2022.

[30] Phong H Nguyen, Kai Xu, Rick Walker, and BL William Wong. Timesets: Timeline visualization with set relations. *Information Visualization*, 15(3):253–269, 2016.

[31] Xavier Ouvrard, Jean-Marie Le Goff, and Stéphane Marchand-Maillet. Networks of collaborations: Hypergraph modeling and visualisation. *arXiv preprint arXiv:1707.00115*, 2017.

[32] Jesse Paquette and Taku Tokuyasu. Hypergraph visualization and enrichment statistics: how the egan paradigm facilitates organic discovery from big data. In *Human Vision and Electronic Imaging XVI*, volume 7865, pages 126–143. SPIE, 2011.

[33] Deokgun Park, Seungyeon Kim, Jurim Lee, Jaegul Choo, Nicholas Diakopoulos, and Niklas Elmqvist. Conceptvector: Text visual analytics via interactive lexicon building using word embedding. *IEEE transactions on visualization and computer graphics*, 24(1):361–370, 2017.

[34] Fernando V Paulovich, Franklina MB Toledo, Guilherme P Telles, Rosane Minghim, and Luis Gustavo Nonato. Semantic wordification of document collections. In *Computer Graphics Forum*, volume 31, pages 1145–1153. Wiley Online Library, 2012.

[35] John Stasko, Carsten Gorg, Zhicheng Liu, and Kanupriya Singhal. Jigsaw: supporting investigative analysis through interactive visualization. In *2007 IEEE Symposium on Visual Analytics Science and Technology*, pages 131–138. IEEE, 2007.

[36] Dirk Streeb, Devanshu Arya, Daniel A Keim, and Marcel Worring. Visual analytics framework for the assessment of temporal hypergraph prediction models. In *Set Visual Analytics Workshop at IEEE VIS 2019*, 2019.

[37] Hai-Long Trieu, Thy Thy Tran, Khoa NA Duong, Anh Nguyen, Makoto Miwa, and Sophia Ananiadou. Deepeventmine: end-to-end neural nested event extraction from biomedical texts. *Bioinformatics*, 36(19):4910–4917, 2020.

[38] Paola Valdivia, Paolo Buono, Catherine Plaisant, Nicole Dufournaud, and Jean-Daniel Fekete. Analyzing dynamic hypergraphs with parallel aggregated ordered hypergraph visualization. *IEEE transactions on visualization and computer graphics*, 27(1):1–13, 2019.

[39] Ike Vayansky and Sathish AP Kumar. A review of topic modeling methods. *Information Systems*, 94:101582, 2020.

[40] Fernanda B Viégas and Martin Wattenberg. Timelines tag clouds and the case for vernac-

ular visualization. *interactions*, 15(4):49–52, 2008.

[41] Ji Wang, Jian Zhao, Sheng Guo, Chris North, and Naren Ramakrishnan. Recloud: semantics-based word cloud visualization of user reviews. In *Graphics Interface 2014*, pages 151–158. AK Peters/CRC Press, 2020.

[42] Yingcai Wu, Thomas Provan, Furu Wei, Shixia Liu, and Kwan-Liu Ma. Semantic-preserving word clouds by seam carving. In *Computer Graphics Forum*, volume 30, pages 741–750. Wiley Online Library, 2011.