# TextCube: Automated Construction and Multidimensional Exploration

Yu Meng Jiaxin Huang Jingbo Shang Jiawei Han
Department of Computer Science, University of Illinois at Urbana-Champaign, Illinois, USA
{yumeng5, jiaxinh3, shang7, hanj}@illinois.edu

## ABSTRACT

Today's society is immersed in a wealth of text data, ranging from news articles, to social media, research literature, medical records, and corporate reports. A grand challenge of data science and engineering is to develop effective and scalable methods to extract structures and knowledge from massive text data to satisfy diverse applications, without extensive, corpus-specific human annotations.

In this tutorial, we show that TextCube provides a critical information organization structure that will satisfy such an information need. We overview a set of recently developed data-driven methods that facilitate automated construction of TextCubes from massive, domain-specific text corpora, and show that TextCubes so constructed will enhance text exploration and analysis for various applications. We focus on new TextCube construction methods that are scalable, weakly-supervised, domain-independent, language-agnostic, and effective (i.e., generating quality TextCubes from large corpora of various domains). We will demonstrate with real datasets (including news articles, scientific publications, and product reviews) on how TextCubes can be constructed to assist multidimensional analysis of massive text corpora.

**PVLDB Reference Format:**
Yu Meng, Jiaxin Huang, Jingbo Shang, Jiawei Han. TextCube: Automated Construction and Multidimensional Exploration. *PVLDB*, 12(12): 1974-1977, 2019.
DOI: https://doi.org/10.14778/3352063.3352113

## 1. INTRODUCTION

In this tutorial we present our vision on TextCube and then introduce three technical modules as outlined below.

### TextCube: An Introduction and General Vision

1. What is TextCube?
2. Why TextCube?
3. TextCube: Existing methods, challenges, new developments, and applications

### Module I. Mining Structural Primitives from Text: Phrases, Entities and Relations

Mining structured factual information is the premise in text analysis to turn unstructured text into textual structures (*e.g.*, entities, concept taxonomies) for TextCube construction. The factual information in massive text corpora usually consists of entity mentions and relations between them, which are typically described by quality phrases. Thus, we first introduce data-driven approaches for mining phrases, entities and relations from unstructured text corpora.

1. **Phrase Mining:** Automated phrase mining methods utilize word occurrence statistics in the corpus and/or incorporate external information in public and general knowledge bases. It can be categorized as (1) unsupervised phrase mining methods (e.g., ToPMine [9]) that evaluate word co-occurrence frequency, collocation and completeness and explore phrasal segmentation; (2) weakly supervised methods (e.g., [19, 15]) that introduce classification to enhance the quality of phrase mining; and (3) distantly supervised methods (e.g., AutoPhrase [37]) that remove the need of manual labeling effort in classification with the assistance of existing knowledge bases (e.g., Wikipedia).
2. **Entity Recognition and Typing:** Entity recognition and typing can be conducted at both corpus level and sentence level. For corpus level entity recognition and typing, entity mentions are extracted by phrase mining algorithms and then linked [40] to external knowledge bases. To determine the types of unlinkable entities, coarse-level typing algorithms (e.g., ClusType [32]) and fine-grained typing algorithms (e.g., PLE [34], AFET [33]) leverage distant supervision from the types of known entities in the knowledge bases. For sentence level entity recognition, sequence labeling methods (e.g., LSTM-CRF [14], CNN-LSTM-CRF [22], and LM-LSTM-CRF [20]) achieve state-of-the-art on benchmark datasets under the supervised setting. Going beyond, a recent named entity recognition method, AutoNER [38], uses only distant supervision from knowledge bases without any line-by-line human annotation and exploits a new labeling scheme to learn a robust named entity tagger that achieves competitive results with supervised methods on benchmark datasets.
3. **Relation Extraction:** Relation extraction methods can be categorized into supervised methods for general domains and weakly-supervised methods for specific domains. In the context of text documents from general

*Proceedings of the VLDB Endowment,* Vol. 12, No. 12
ISSN 2150-8097.
DOI: https://doi.org/10.14778/3352063.3352113

domain (*e.g.*, news), we introduce supervised relation extraction systems, which are trained on large amount of human-annotated data and rely on the output of entity mention detectors. We discuss kernel-based supervised approaches [30], examine different features used in the systems [57], and introduce recent popular neural network models [10]. In context of domain-specific text corpora, we focus on domain-independent methods that rely on minimal human-annotated data: Weak supervision [31] or distant supervision (with the help of public knowledge bases) [50, 35]. Finally, we present ReMine [58] which extracts high-confidence relational phrases from domain-specific texts in an end-to-end manner.

### Module II. Automated Construction of TextCubes

With a brief introduction to data cube and OLAP (online analytical processing) research in database systems, we introduce the TextCube model and its construction methods by first discussing taxonomy construction, then reviewing word embedding learning models, and presenting recent TextCube construction techniques.

1. **Taxonomy construction:** Taxonomy construction clusters similar concepts and generates a hierarchy of "concept clusters" from massive corpus. We first introduce a line of work that constructs cluster-based taxonomy based on hierarchical topic models [3, 29, 2], then discuss methods that use more general probabilistic graphical models, including Bayesian Rose Tree [21, 41], Sparse Backoff Tree [8], and phrase-centric models [47, 48, 49]. Most of these techniques are based on one-hot representations and implicitly model word granularity by making some statistical assumptions. Finally, we present TaxoGen [53], a recursive framework that leverages word distributional representations and constructs cluster-based taxonomy using adaptive spherical clustering and local embedding.
2. **Embedding learning:** We introduce a few word embedding learning frameworks since they serve as the preliminary to document classification and TextCube construction. These include the well-known unsupervised word embedding models, such as Word2Vec [28, 27], which preserves local context similarity of words for embedding learning, and several other word embedding frameworks that incorporate other types of context in word embedding, such as global contexts [11]. We also discuss a recent popular deep language model BERT [6] that learns text representations via bidirectional deep transformers guided by a masked language model objective. Finally, we introduce JoSE, an unsupervised text embedding framework that jointly learns word embedding and paragraph embedding by incorporating both local and global contexts to capture more complete text semantics, and present TopicMine [24], a category-name guided word embedding framework that endows word embedding with discriminative power over the specific set of categories.
3. **Supervised methods:** We introduce a variety of supervised methods for text cube construction. Relying on a sufficient number of document-label training pairs, these methods learn reliable classifiers that are capable of predicting the label of a new document, including support vector machines [12], decision trees [1, 36], and neural networks [51]. We present how to adapt the supervised methods for text cube construction along with their strength and drawbacks.
4. **Weakly-supervised methods:** Due to the high cost of providing sufficient, high-quality document-label pairs for supervision, weakly-supervised or unsupervised methods for text categorization are in great need. We introduce methods using heuristic rules [13] to generate training data, as well as methods leveraging external knowledge such as *Wikipedia* [5, 42] to construct text cubes. After that, we present Doc2Cube [44], a joint embedding framework that allocates documents into cube cells by computing dimension-aware document representation and iteratively expanding from initial class label surface names to incorporate more high-quality class-distinctive phrases. Finally, we present WeSTClass [25] and WeSHClass [26], which generate pseudo training data for neural classifier pre-training, and then bootstrap the classifier by self-training on unlabeled documents.

### Module III. Multi-Dimensional Exploration of TextCubes

TextCube facilitates multidimensional text analysis. The techniques that leverage the TextCube structure for multi-dimensional text exploration are introduced here.

1. **Cube-based multidimensional analysis:** With TextCube, online analytical processing (OLAP) such as drill-down and roll-up can be performed for efficient searching, querying and retrieval of text data. We present methods that perform multidimensional analysis based on the TextCube structure, including efficient aggregated measure computation [18, 23], cube cell summarization [43], and keyword-based interactive exploration [7, 56, 54]. CASeOLAP [45] is introduced here, which conducts comparative analysis over sibling cells to extract top representative phrases for each cube cell.
2. **Text summarization:** Another important application of TextCube is the summarization of documents belonging to a specific set of categories. For example, one can analyze the set of documents in the cube cell "⟨"*US*", "*Gun Control*"⟩" to extract top-$k$ frequent and discriminative phrases dedicated to the issues related to "gun control in the US" by comparing the cell with its siblings and parents/children. We will introduce first a line of extractive summarization methods that select text pieces from sources and arrange them to form the summary, including clustering-based methods [46, 4, 55] and graph-based methods [39, 52], and then some recent neural-model-based abstractive summarization frameworks that generate summarizations. The neural models are typically trained to first model latent semantics of sentences via an encoder, and then generate summarizations by decoding hidden states of the latent space [16, 17].

Our tutorial also includes a system demonstration that shows the capabilities of the tools and methods mentioned on a variety of test cases and metrics, including a few case-studies on real-world datasets consisting of news articles and scientific literature. The tutorial will be concluded by a discussion on potential applications, related tasks, and future research directions.

**Targeted Audience and Assumed Background.** The audience with a good background in database systems, data mining, text mining, natural language processing, and machine learning, will benefit most from this tutorial. However, we believe the tutorial would give general audience and newcomers an introductory pointer to the current work and important research topics on this theme, and inspire them to learn more. Only preliminary knowledge about text mining, data mining, and their applications are needed.

## 2. BIOGRAPHIES OF PRESENTERS

- **Yu Meng:** Ph.D. student, Computer Science, UIUC. His research focuses on mining structured knowledge from massive text corpora with minimum human supervision.
- **Jiaxin Huang:** Ph.D. student, Computer Science, UIUC. Her research focuses on mining structured knowledge from massive text corpora. She is the recipient of Chirag Foundation Graduate Fellowship in Computer Science.
- **Jingbo Shang:** Ph.D. candidate, Computer Science, UIUC. His research focuses on mining and constructing structured knowledge from massive text corpora with minimum human effort. His research has been recognized with multiple prestigious awards, including Grand Prize of Yelp Dataset Challenge in 2015, Google PhD Fellowship in Structured Data and Database Management in 2017. He has rich experiences in delivering tutorials in major conferences (SIGMOD'17, WWW'17, SIGKDD'17, and SIGKDD'18).
- **Jiawei Han:** Abel Bliss Professor, Computer Science, UIUC. His research areas encompass data mining, text mining data warehousing and information network analysis, with over 800 research publications. He is Fellow of ACM, Fellow of IEEE, and received numerous prominent awards, including ACM SIGKDD Innovation Award (2004) and IEEE Computer Society W. Wallace McDowell Award (2009). He has delivered 50+ conference tutorials or keynote speeches (*e.g.*, KDD 2018 tutorial and WSDM 2018 keynote).

## 3. ACKNOWLEDGEMENTS

This research is sponsored in part by U.S. Army Research Lab. under Cooperative Agreement No. W911NF-09-2-0053 (NSCTA), DARPA under Agreement No. W911NF-17-C-0099, National Science Foundation IIS 16-18481, IIS 17-04532, and IIS-17-41317, DTRA HDTRA11810026, and grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov).

## 4. REFERENCES

[1] C. C. Aggarwal and C. Zhai. A survey of text classification algorithms. *Mining text data*, pages 163–222, 2012.

[2] D. M. Blei, T. L. Griffiths, and M. I. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of ACM*, 2010.

[3] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *NIPS*, 2003.

[4] X. Cai and W. Li. Ranking through clustering: An integrated approach to multi-document summarization. *IEEE Transactions on Audio, Speech, and Language Processing*, 21:1424–1433, 2013.

[5] X. Chen, Y. Xia, P. Jin, and J. A. Carroll. Dataless text classification with descriptive LDA. In *AAAI*, pages 2224–2231, 2015.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

[7] B. Ding, B. Zhao, C. X. Lin, J. Han, and C. Zhai. Topcells: Keyword-based search of top-k aggregated documents in text cube. In *ICDE*, pages 381–384, 2010.

[8] D. Downey, C. Bhagavatula, and Y. Yang. Efficient methods for inferring large sparse topic hierarchies. In *ACL*, 2015.

[9] A. El-Kishky, Y. Song, C. Wang, C. R. Voss, and J. Han. Scalable topical phrase mining from text corpora. *PVLDB*, 8(3):305–316, 2014.

[10] M. R. Gormley, M. Yu, and M. Dredze. Improved relation extraction with feature-rich compositional embedding models. *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2015.

[11] E. H. Huang, R. Socher, C. D. Manning, and A. Y. Ng. Improving word representations via global context and multiple word prototypes. In *ACL*, 2012.

[12] T. Joachims. Transductive inference for text classification using support vector machines. In *ICML*, pages 200–209, 1999.

[13] Y. Ko and J. Seo. Automatic text categorization by unsupervised learning. In *COLING*, pages 453–459, 2000.

[14] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*, 2016.

[15] B. Li, X. Yang, B. Wang, and W. Cui. Efficiently mining high quality phrases from texts. In *AAAI*, pages 3474–3481, 2017.

[16] C. Li, W. Xu, S. Li, and S. Gao. Guiding generation for abstractive text summarization based on key information guide network. In *NAACL-HLT*, 2018.

[17] P. Li, Z. Wang, W. Lam, Z. Ren, and L. Bing. Salience estimation via variational auto-encoders for multi-document summarization. In *AAAI*, 2017.

[18] C. X. Lin, B. Ding, J. Han, F. Zhu, and B. Zhao. Text cube: Computing IR measures for multidimensional text database analysis. In *ICDM*, pages 905–910, 2008.

[19] J. Liu, J. Shang, C. Wang, X. Ren, and J. Han. Mining quality phrases from massive text corpora. In *SIGMOD*, 2015.

[20] L. Liu, J. Shang, F. Xu, X. Ren, H. Gui, J. Peng, and J. Han. Empower sequence labeling with task-aware neural language model. *arXiv preprint arXiv:1709.04109*, 2017.

[21] X. Liu, Y. Song, S. Liu, and H. Wang. Automatic taxonomy construction from keywords. In *KDD*, 2012.

[22] X. Ma and E. Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*, 2016.

[23] M. Mendoza, E. Alegría, M. Maca, C. A. C. Lozada, and E. León. Multidimensional analysis model for a document warehouse that includes textual measures. *Decision Support Systems*, 72:44–59, 2015.

[24] Y. Meng, J. Huang, Z. Wang, C. Fan, G. Wang, C. Zhang, J. Shang, L. Kaplan, and J. Han. Topicmine: User-guided topic mining by category-oriented embedding. 2019.

[25] Y. Meng, J. Shen, C. Zhang, and J. Han. Weakly-supervised neural text classification. In *CIKM*, 2018.

[26] Y. Meng, J. Shen, C. Zhang, and J. Han. Weakly-supervised hierarchical text classification. In *AAAI*, 2019.

[27] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[28] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

[29] D. M. Mimno, W. Li, and A. D. McCallum. Mixtures of hierarchical topics with pachinko allocation. In *ICML*, 2007.

[30] R. J. Mooney and R. C. Bunescu. Subsequence kernels for relation extraction. In *NIPS*, 2005.

[31] M. Qu, X. Ren, Y. L. Zhang, and J. Han. Weakly-supervised relation extraction by pattern-enhanced embedding learning. In *WWW*, 2018.

[32] X. Ren, A. El-Kishky, C. Wang, F. Tao, C. R. Voss, and J. Han. Clustype: Effective entity recognition and typing by relation phrase-based clustering. In *SIGKDD*, 2015.

[33] X. Ren, W. He, M. Qu, L. Huang, H. Ji, and J. Han. Afet: Automatic fine-grained entity typing by hierarchical partial-label embedding. In *EMNLP*, 2016.

[34] X. Ren, W. He, M. Qu, C. R. Voss, H. Ji, and J. Han. Label noise reduction in entity typing by heterogeneous partial-label embedding. In *SIGKDD*, 2016.

[35] X. Ren, Z. Wu, W. He, M. Qu, C. R. Voss, H. Ji, T. F. Abdelzaher, and J. Han. CoType: Joint extraction of typed entities and relations with knowledge bases. In *WWW*, 2017.

[36] F. Sebastiani. Machine learning in automated text categorization. *ACM computing surveys*, 34(1):1–47, 2002.

[37] J. Shang, J. Liu, M. Jiang, X. Ren, C. R. Voss, and J. Han. Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering*, 30:1825–1837, 2018.

[38] J. Shang, L. Liu, X. Gu, X. Ren, T. Ren, and J. Han. Learning named entity tagger using domain-specific dictionary. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2054–2064, 2018.

[39] C. Shen and T. Li. Multi-document summarization via the minimum dominating set. In *COLING*, 2010.

[40] W. Shen, J. Han, and J. Wang. A probabilistic model for linking named entities in web text with heterogeneous information networks. In *SIGMOD Conference*, 2014.

[41] Y. Song, S. Liu, X. Liu, and H. Wang. Automatic taxonomy construction from keywords via scalable bayesian rose trees. *TKDE*, 2015.

[42] Y. Song and D. Roth. On dataless hierarchical text classification. In *AAAI*, pages 1579–1585, 2014.

[43] F. Tao, K. H. Lei, J. Han, C. Zhai, X. Cheng, M. Danilevsky, N. Desai, B. Ding, J. Ge, H. Ji, R. Kanade, A. Kao, Q. Li, Y. Li, C. X. Lin, J. Liu, N. C. Oza, A. N. Srivastava, R. Tjoelker, C. Wang, D. Zhang, and B. Zhao. Eventcube: multi-dimensional search and mining of structured and text data. In *KDD*, pages 1494–1497, 2013.

[44] F. Tao, C. Zhang, X. Chen, M. Jiang, T. Hanratty, L. M. Kaplan, and J. Han. Doc2cube: Allocating documents to text cube without labeled data. *2018 IEEE International Conference on Data Mining (ICDM)*, pages 1260–1265, 2018.

[45] F. Tao, H. Zhuang, C. W. Yu, Q. Wang, T. Cassidy, L. R. Kaplan, C. R. Voss, and J. Han. Multi-dimensional, phrase-based summarization in text cubes. *IEEE Data Eng. Bull.*, 39(3):74–84, 2016.

[46] X. Wan and J. Yang. Multi-document summarization using cluster-based link analysis. In *SIGIR*, 2008.

[47] C. Wang, M. Danilevsky, N. Desai, Y. Zhang, P. Nguyen, T. Taula, and J. Han. A phrase mining framework for recursive construction of a topical hierarchy. In *KDD*, 2013.

[48] C. Wang, J. Liu, N. Desai, M. Danilevsky, and J. Han. Constructing topical hierarchies in heterogeneous information networks. *ICDM*, 2013.

[49] C. Wang, X. Liu, Y. Song, and J. Han. Towards interactive construction of topical hierarchy: A recursive tensor decomposition approach. *KDD*, 2015.

[50] Z. Wu, X. Ren, F. F. Xu, J. Li, and J. Han. Indirect supervision for relation extraction using question-answer pairs. In *WSDM*, 2018.

[51] Z. Yang, D. Yang, C. Dyer, X. He, A. J. Smola, and E. H. Hovy. Hierarchical attention networks for document classification. In *HLT-NAACL*, pages 1480–1489, 2016.

[52] M. Yasunaga, R. Zhang, K. Meelu, A. Pareek, K. Srinivasan, and D. R. Radev. Graph-based neural multi-document summarization. In *CoNLL*, 2017.

[53] C. Zhang, F. Tao, X. Chen, J. Shen, M. Jiang, B. M. Sadler, M. T. Vanni, and J. Han. Taxogen: Constructing topical concept taxonomy by adaptive term embedding and clustering. In *KDD 2018*, 2018.

[54] D. Zhang, C. Zhai, and J. Han. Topic cube: Topic modeling for olap on multidimensional text databases. In *SDM*, 2009.

[55] Y. Zhang, Y. Xia, Y. Liu, and W. Wang. Clustering sentences with density peaks for multi-document summarization. In *HLT-NAACL*, 2015.

[56] B. Zhao, C. X. Lin, B. Ding, and J. Han. Texplorer: keyword-based object search and exploration in multidimensional text databases. In *CIKM*, pages 1709–1718, 2011.

[57] G. Zhou, J. Su, J. Zhang, and M. Zhang. Exploring various knowledge in relation extraction. In *ACL*, 2005.

[58] Q. Zhu, X. Ren, J. Shang, Y. Zhang, A. El-Kishky, and J. Han. Integrating local context and global cohesiveness for open information extraction. In *WSDM*, 2019.