TopicMine: User-Guided Topic Mining by Category-Oriented Embedding

Yu Meng[1*], Jiaxin Huang[1*], Zihan Wang[1], Chenyu Fan[1], Guangyuan Wang[1], Chao Zhang[2], Jingbo Shang[1], Lance Kaplan[3], Jiawei Han[1]
[1]University of Illinois at Urbana-Champaign [2]Georgia Institute of Technology [3]U.S. Army Research Laboratory
[1]{yumeng5, jiaxinh3, zihanw2, chenyuf2, gwang10, shang7, hanj}@illinois.edu
[2]chaozhang@gatech.edu [3]lance.m.kaplan.civ@mail.mil

ABSTRACT

With an ever-increasing volume of textual data coming from news reports, social media, literature articles, and medical records, it becomes a necessity to distill knowledge from text data by categories according to users' interests. For example, given a general news corpus, one user may be interested in organizing articles by countries; whereas another may want to browse articles by themes. In either case, a user's interest can be easily described by a set of category names. In this project, we develop a framework, TopicMine, which takes user-provided category names as guidance and mines *category representative phrases* to form coherent topics. Specifically, TopicMine first leverages a phrase mining tool to extract quality phrases from the text corpus, and then learns an embedding space that best separates the categories specified by the user. Finally, category representative phrases are retrieved by considering both topic relevance and semantic generality. The mined topics identified by category representative phrases facilitate effective and efficient understanding, organizing, searching, and summarizing of textual contents based on users' needs.

CCS CONCEPTS

• **Information systems → Users and interactive retrieval**; *Retrieval models and ranking*;

KEYWORDS

Topic mining, key phrase retrieval, word embedding

1 INTRODUCTION

Topic discovery in massive text corpora, presenting a holistic view of the contents to users, has long been a keystone in text understanding. Conventional topic modeling approaches like Latent Dirichlet Allocation (LDA) [1] performs topic discovery in a purely unsupervised manner, by modeling the text generation process. However, the topics so discovered may not be completely interpretable, or are not of the user's interests [2], which hinders their practical applicability. Some later studies such as [3] extend unsupervised topic modeling by incorporating user-provided seed-words, but they still suffer from the following limitations: (1) The text corpus is modeled with bag-of-words assumption without exploiting word-order information and semantic similarity between words. (2) They leverage must-link and cannot-link relationships between seed words within the same topic and across different topics, and therefore require users to provide multiple seed words for every topic, which can be difficult for non-experts. Therefore, it is of great importance to develop a system that mines coherent topics that meet users' needs, with minimal user supervision.

Research Project and Goals. In this project, we develop TopicMine, a user-guided topic mining system that retrieves *category representative phrases* from massive text corpora based on user-provided category names. To achieve this goal, TopicMine first mines quality phrases from the target corpus, and learns an embedding space that best separates the user-interested categories based on the provided category names. Moreover, phrase semantic specificity is trained jointly with phrase embedding, so that category representative phrases are not only retrieved by their relevance to the categories, but also sorted by their semantic generality, which helps users understand the mined topics in a coarse-to-fine manner.

Fit with the KDD Ecosystem. Our system effectively retrieves category representative phrases from large-scale text corpora to form coherent topics based on user's interests, which facilitates many fundamental tasks in Data Mining, Information Retrieval, Natural Language Processing, and Machine Learning, such as topic modeling, text classification, and document retrieval. Our project is also easily accessible to general audiences, as our system only requires users to provide their own corpus and category names without need for further data mining or machine learning knowledge, which is user-friendly for non-experts.

2 INNOVATION

In this section, we introduce the major innovations in our system.

Incorporating Local and Global Textual Contexts in Embedding Learning. Traditional topic modeling and word embedding frameworks model text in complementary perspectives: Topic modeling leverages word-document co-occurrence statistics to derive document-topic distributions and topic-word distributions, while word embedding learns word representation from word-word co-occurrence in local context windows. The use of local context (surrounding words) and global context (belonging documents) respectively has shown great effectiveness in our previous works [4, 6], which encourages us to incorporate both local context and global context into phrase embedding learning of TopicMine, so that the

*Equal Contribution.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Table 1: Topic mining results on New York Times dataset (NYT) and Yelp Review dataset (Yelp). The table header denotes category types and corresponding category names as input to TopicMine**.**

NYT-Location			**Yelp**-Food			**Yelp**-Sentiment	
britain	china	canada	steak	seafood	burger	good	bad
england	beijing	ontario	sirloin steak	oysters	burgers	great	sucky
london	shanghai	toronto	hanger steak	mussels	cheeseburger	delicious	sickening
britons	hong kong	quebec	chicken fried steak	clams	hamburger	mindful	nasty
scottish	fujian	montreal	skirt steak	anchovies	burger king	excellent	dreadful
great britain	hubei	ottawa	flank steak	tilapia	hamburgers	wonderful	freaks
british government	nanjing	alberta	striploin	monkfish	smash burger	faithful	cheapskates
wales	liaoning	vancouver	roast beef	shellfish	whoppers	keen	snot
scotland	guangxi	calgary	roast pork	sardines	in n out burger	inspiring	horrible
united kingdom	anhui	manitoba	sirloin	seared scallops	patty melt	courteous	misery
yorkshire	hangzhou	british columbia	corned beef	seared tuna	smash fries	wholesome	stinks

Table 2: Phrases of category "Science" from different ranges of distributional specificity.

Range of κ	Science ($\kappa_{\text{science}} = 0.539$)
$0.539 < \kappa < 0.6$	scientist, academic, research, laboratory
$0.6 < \kappa < 0.8$	physics, sociology, biology, astronomy
$0.8 < \kappa < 1.0$	microbiology, anthropolgy, physiology, cosmology
$\kappa > 1.0$	national science foundation, george washington university, hong kong university, american academy

semantics of phrases and category-distinctive information can be more effectively encoded into the learned embedding.

Weakly-Supervised Phrase Embedding. We first use a phrase mining pipeline [5] to extract quality phrases from the text corpus. Then, to learn a phrase embedding space where the user-interested categories are best separated, TopicMine explicitly models category semantics by jointly learning category embedding and word embedding, with a regularization term that enforces category representative words to be embedded close to their corresponding category and far from other categories in the joint embedding space. During embedding training, category representative phrases are gradually discovered based on category relevance and contribute to the regularization of the embedding space so that the category embedding manifold reflects more and more complete category semantics.

Distributional Specificity and Topic Presentation. Traditional topic modeling presents a category via a top ranked list according to topic-word distribution in each category, which usually seems randomly-ordered because latent probability distribution is generally hard to be interpreted. In TopicMine, we propose a way to present a category in a more user-friendly way: The category representative phrases are first selected by category relevance, and then ranked by semantic specificity in a coarse-to-fine manner. For example, "Texas" will be ranked higher than "Austin" as representative phrases for category "The United States". To achieve this property, TopicMine learns phrase distributed specificity during the embedding learning process: We learn an additional parameter κ for each phrase which reflects how specific the phrase meaning is based on how variant the phrase's local contexts are in the entire corpus. The user can also set threshold values to obtain filtered phrases within a specific range of distributional specificity, as shown in Table 2.

3 DEMONSTRATION

We have developed a web-based system for researchers and practitioners to easily interact with TopicMine. We show the system outputs on several types of categories from two real-world datasets in Table 1. The demonstration video and final system will be gradually rolled out at http://dmserv4.cs.illinois.edu/topicmine_demo/.

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).

REFERENCES

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. In *NIPS*.
[2] Jonathan Chang, Jordan L. Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In *NIPS*.
[3] Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. 2012. Incorporating lexical priors into topic models. In *EACL*.
[4] Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. 2019. Weakly-Supervised Hierarchical Text Classification. In *AAAI*.
[5] Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R. Voss, and Jiawei Han. 2018. Automated Phrase Mining from Massive Text Corpora. *IEEE Transactions on Knowledge and Data Engineering* 30 (2018), 1825–1837.
[6] Fangbo Tao, Chao Zhang, Xiusi Chen, Meng Jiang, Tim Hanratty, Lance M. Kaplan, and Jiawei Han. 2018. Doc2Cube: Allocating Documents to Text Cube Without Labeled Data. In *ICDM*.