FTS: Faceted Taxonomy Construction and Search for Scientific Publications

Hanwen Zha[1] Jiaming Shen[2] Keqian Li[1] Warren Greiff[3]
Michelle T. Vanni[4] Jiawei Han[2] Xifeng Yan[1]
[1]University of California, Santa Barbara [2]University of Illinois at Urbana-Champaign
[3]MITRE Corporation [4]U.S. Army Research Laboratory
[1]{hwzha, klee, xyan}@cs.ucsb.edu [2]{js2, hanj}@illinois.edu [3]greiff@mitre.org [4]michelle.t.vanni.civ@mail.mil

ABSTRACT

Scalable keyword-based information retrieval has dominated the search industry for decades. When performing a sophisticated intelligence search and analysis task, a user is challenged to pose a right query, read multiple retrieved articles, understand their major contents, discover more relevant terms, and iterate. This process is often ad hoc and in many cases, very challenging especially when researchers start to explore a field they are not familiar with. For tasks like summarizing research efforts in one area, an analyst needs to interact with a keyword-based search engine for a long time before a reasonable, comprehensive technical report can be written. In this work, we developed a network-based, unified search and navigation platform, called **FTS** (Faceted Taxonomy Construction and Search), to ease query development and facilitate intelligence exploration in a large text repository, focused on scientific publications. It leverages the newest phrase mining, concept embedding and deep learning techniques to automatically extract concept terms and link them in a taxonomy structure, which could facilitate many interesting downstream applications including summarization, trend analysis, document categorization and recommendation.

KEYWORDS

Phrase Mining; Taxonomy; Text Categorization; Document Search

1 INTRODUCTION

The number of scientific publications is ever increasing. According to the prominent STM report, the number of journal articles published in 2014 alone approached 2.5 million. It takes much longer time to digest a scientific paper than a webpage, posting a hard constraint on the number of papers a researcher can read. The problem becomes much severe for intelligence analysts who need to browse papers and quickly grasp the major activities in new research areas. They either rely on manually constructed taxonomies (e.g. ACM Computing Classification System) or figure it out through their own reading. Manually constructing taxonomies requires a large amount of human effort, which is not only expensive, but also could be quickly outdated for fast developing areas. Researchers or analysts thus have to read papers and blogs by themselves for finding right keywords to keep track of emerging topics and have a comprehensive view of a research area.

In order to address the aforementioned issues, we develop **FTS** to automatically mine concept terms from a technical corpus selected by the user and construct faceted taxonomy for fast exploration of scientific publications. FTS supports multiple functionalities, including taxonomy construction, unsupervised document categorization, query suggestion and trend analysis, to save the literature search and analysis time.

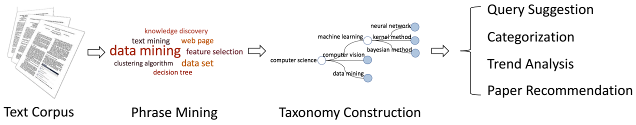


Figure 1: The Workflow of FTS

In FTS, we re-examined the automatic taxonomy construction problem [3] by adopting the newest embedding and deep learning techniques. FTS was built upon our latest research results in text mining [1, 2, 4–7], many of which were published in the recent SIGKDD conferences. The FTS project was developed by researchers from the Army Research Lab, MITRE, UCSB and UIUC.

2 MAIN FUNCTIONS

The workflow of FTS system is depicted in Figure 1. It has the following modules.

2.1 Phrase Mining

FTS adopts advanced phrase mining techniques such as SegPhrase [2] and AutoPhrase [4] developed by our team to mine high-quality phrases from massive text data.

The main idea behind these data-driven approaches is to find frequent n-grams from text and rectify the raw frequency with some quality estimation criteria. A classifier is trained to classify phrases based on quality estimation features and labeled phrase examples. Furthermore, with distant supervision from general knowledge bases such as Wikipedia, training labels are automatically generated without human effort. FTS analyzed the title and abstract of 1.2 million computer science papers downloaded from *DBLP*[1] and *Semantic Scholar*[2] and extracted around 180k concept terms.

There are very frequent concepts such as “machine learning” and “data mining,” as well as rare, but interesting ones such as “quantum learning” and “quantum neural networks.” The mined concepts

KDD'18, August 19-23, 2018, London, UK
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5552-0/18/08…$15.00

[1] https://dblp.uni-trier.de/

[2] https://www.semanticscholar.org/

are building blocks for downstream modules such as taxonomy construction and trend analysis. The text and meta information, as well as the mined concept terms, are indexed by Elasticsearch for fast document retrieval.

2.2 Taxonomy Construction

Among the mined concept terms, the important ones are selected and connected together to generate a taxonomy from the underlying text corpus. We have two recent works, TaxonGen [7] and HiExpan [6], to be published in SIGKDD 2018. Both of them developed sophisticated taxonomy construction algorithms.

TaxonGen [7] generates taxonomy with hierarchical clustering and term embeddings. Each node of the resulting taxonomy is a cluster of concept terms and its representative term is automatically selected. The spherical clustering and locally-trained term embeddings were proposed to boost the taxonomy quality. HiExpan [6] focuses on interactive taxonomy generation under user guidance. With weakly-supervised set expansion and depth expansion, the taxonomy is constructed to fit the user's desire. The generated results could be further adapted to multi-faceted taxonomies with different facets. With AutoPhrase, TaxonGen and HiExpan, FTS can dynamically generate concept terms and taxonomies based on a subset of documents identified by users. Users can also interact with the system to further refine the taxonomy.

Suppose an analyst wants to investigate what is going on in an emerging research area, e.g., "quantum learning." She could either upload a paper collection related to "quantum learning" or use Elasticsearch to retrieve a set of related papers in our system. Next, she could generate a taxonomy for this specific collection. Based on the result, she may modify the paper collection or query (e.g. adding an additional term "quantum neural networks"), repeat the above process and refine the taxonomy. These two steps can iterate.

2.3 Intelligent Categorization and Search

Once a taxonomy is built, the next step is to put publications in different taxonomy nodes. An unsupervised document categorization technique, UNEC [1], was developed in FTS. UNEC is a cascade embedding approach: Based on a concept similarity graph built from concept embedding, the concepts are embedded into a hidden category space given only category names. UNEC can quickly help analysts identify research hot spots in a taxonomy.

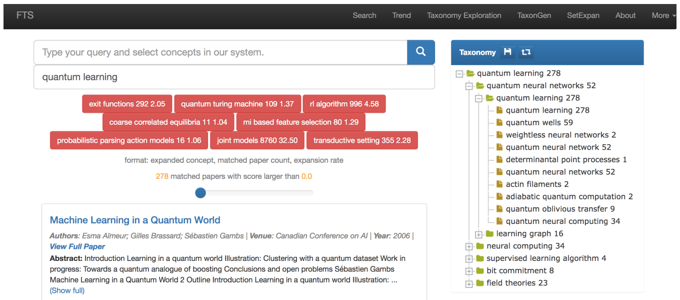
In addition to categorization, FTS also provides query suggestion capability which can suggest terms related to a user query. This function can help a user to adjust the targeted documents for further analysis. For example, given a query "quantum learning," FTS could suggest "quantum Turing machine," "RL algorithm" etc. FTS developed an embedding based query suggestion approach, which leverages the mined concepts and discovers the relation among these concepts through word embedding. In addition, SetExpan [5] is adopted in FTS in order to expand the query set and rank related concepts. SetExpan expands the query set by using skip-gram features and ensembles the ranks of expanded terms.

2.4 Trend Analysis

When the time information of publications is plugged in, FTS can naturally support trend analysis. It could show the research focus change over time, and compare the research strength of different countries or topics, e.g. showing the strength and the developing trend of "quantum learning" in countries like Russia, China, and the United States. In that way, analysts can be aware of subareas that are booming as well as who are actively involved in those areas.

3 DEMONSTRATION

We developed a user-friendly web site for analysts to easily interact with FTS. A prototype system is available at http://fts.cs.ucsb.edu/. Figure 2 shows a screen shot of FTS.



Type your query and select concepts in our system.

quantum learning

exit functions 292 2.05 | quantum turing machine 109 1.37 | rl algorithm 996 4.58 | coarse correlated equilibria 11 1.04 | mi based feature selection 80 1.29 | probabilistic parsing action models 16 1.06 | joint models 8760 32.50 | transductive setting 355 2.28

format: expanded concept, matched paper count, expansion rate

278 matched papers with score larger than 0.0

Machine Learning in a Quantum World

Authors: Esma Aïmeur; Gilles Brassard; Sébastien Gambs | *Venue: Canadian Conference on AI* | *Year: 2006* |
View Full Paper

Abstract: Introduction Learning in a quantum world Illustration: Clustering with a quantum dataset Work in progress: Towards a quantum analogue of boosting Conclusions and open problems Sébastien Gambs Machine Learning in a Quantum World 2 Outline Introduction Learning in a quantum world Illustration: ...
(Show full)

- quantum learning 278
  - quantum neural networks 52
    - quantum learning 278
      - quantum learning 278
      - quantum wells 59
      - weightless neural networks 2
      - quantum neural network 52
      - determinantal point processes 1
      - quantum neural networks 52
      - actin filaments 2
      - adiabatic quantum computation 2
      - quantum oblivious transfer 9
      - quantum neural computing 34
    - learning graph 16
  - neural computing 34
  - supervised learning algorithm 4
  - bit commitment 8
  - field theories 23

Figure 2: A Snapshot for Query "Quantum Learning": Suggested query terms, query result and generated taxonomy.

In this demo, users can search papers with their own query, interactively refine queries, select targeted documents and generate taxonomies. Users can navigate the taxonomy and pick up a few interesting sub-areas and their papers for detailed examination. This will liberate them from coming up with concept terms they are not familiar with. Users can also visualize trends of different topics through time.

A strong motivation for this demonstration is to show the power of text data mining for analyzing scientific publications. Without FTS, one has to read and manually label many research papers before a comprehensive view can be derived. FTS simplifies this process. We hope FTS will benefit many researchers and promote research towards advanced mining of scientific literature.

ACKNOWLEDGMENTS

This research was sponsored in part by the Army Research Laboratory under cooperative agreements W911NF09-2-0053 and NSF 0954125.

REFERENCES

[1] K. Li, H. Zha, Y. Su, and X. Yan. 2018. Unsupervised Neural Categorization for Scientific Publications *(SDM '18)*.

[2] J. Liu, J. Shang, C. Wang, X. Ren, and J. Han. 2015. Mining Quality Phrases from Massive Text Corpora *(SIGMOD'15)*.

[3] X. Liu, Y. Song, S. Liu, and H. Wang. 2012. Automatic Taxonomy Construction from Keywords *(KDD'12)*.

[4] J. Shang, J. Liu, M. Jiang, X. Ren, C. R. Voss, and J. Han. 2018. Automated Phrase Mining from Massive Text Corpora *(TKDE'18)*.

[5] J. Shen, Z. Wu, D. Lei, J. Shang, X. Ren, and J. Han. 2017. SetExpan: Corpus-Based Set Expansion via Context Feature Selection and Rank Ensemble *(ECML PKDD' 17)*.

[6] J. Shen, Z. Wu, D. Lei, C. Zhang, X. Ren, M. T. Vanni, B. Sadler, and J. Han. 2018. HiExpan: Task-Guided Taxonomy Construction by Hierarchical Tree Expansion *(KDD '18)*.

[7] C. Zhang, F. Tao, X. Chen, J. Shen, M. Jiang, B. Sadler, M. T. Vanni, and J. Han. 2018. TaxoGen: Constructing Topical Concept Taxonomy by Adaptive Term Embedding and Clustering *(KDD '18)*.