Identifying and Understanding Business Trends using Topic Models with Word Embedding

Yun Ning Pek
Engineering Systems and Design Pillar
Singapore University of Technology and Design
8 Somapah Rd, Singapore 487372
yunning_pek@alumni.sutd.edu.sg

Kwan Hui Lim
Information Systems Technology and Design Pillar
Singapore University of Technology and Design
8 Somapah Rd, Singapore 487372
kwanhui_lim@sutd.edu.sg

***Abstract*—Topic modelling and trend analysis are increasingly important in today's digital world, especially for identifying promising business ideas and trends. With the increasing amount of data being generated daily, a key challenge is to effectively identify emerging business ideas/topics and trends from this large volume of data. Towards this effort, we introduce a framework that allows us to identify promising business ideas from a large stream of academic papers. Academic papers are suitable for this purpose as they study emerging areas and problems in different domains. Our framework comprises three main components, namely: (i) a data collection component that retrieves academic papers and their meta-data; (ii) a topic modelling algorithm that combines traditional topic modelling techniques with recent advances in word embeddings; and (iii) a trend analysis component that allows us to visualize the popularity of different business trends/topics across time. Results on a corpus of 237k academic papers show that our proposed methods outperform the standard baselines based on topic coherence scores and also allows us to understand key temporal trends.**

***Index Terms*—Topic Models; Word Embedding; Trend Analysis; Academic Papers**

I. Introduction

With today's highly digitalized world, there is an ever increasing amount of data being generated on a daily basis. Similarly, there are approximately 2.5 million academic papers being published annually [1]. These papers typically investigate important research problems in various domains and provide a overall trend on the emerging areas in different domains. Thus, this large amount of data provides a rich source for identifying promising business ideas and trends before they become mainstream. However, this information overload makes it a challenge to sieve through the large amount of information.

To solve this problem, we develop a framework that collects academic papers over multiple timeframes, identifies important topics and visualizes how these trends change over time. This framework is able to identify promising ideas from a large stream of academic papers, by applying a novel topic modelling technique that combines the traditional Latent Dirichlet Allocation (LDA) algorithm [2] with a word expansion strategy based on various embedding and expansion schemes, such as FastText [3], Word2Vec [4], Doc2Vec [5] and CorEx [6]. Thereafter, we identify the trending business ideas based on the number of academic papers using the representative keywords for each topic.

978-1-7281-0858-2/19/$31.00

II. Proposed Framework

Figure 1 shows an overview of our proposed framework for identifying business ideas/topics and trends from academic papers. Our proposed framework comprises the following three main components:

1) **Data Collection Component**. Our main source of data are academic papers, which represent the leading edge of technology development and ideas. This component uses the Semantic Scholar API to download academic papers and their associated meta-data, such as the paper title, authors, affiliation, abstract, published date, etc. This meta-data is then passed to the next topic modelling component for identifying the promising business ideas/topics from this corpus.
2) **Topic Modelling**. We propose various algorithms that extend the standard LDA algorithm by performing a word expansion on the initial set of keywords generated by LDA (seed keywords). Using these seed keywords, we expand it by n words based on the n most similar words by comparing their word embedding with the seed keyword. We utilize word expansion based on CorEx [6] and various embedding like FastText [3] and Word2Vec [4].
3) **Trend Analysis**. Using the output from Step 2 above, we then detect the changes in number of papers related to the identified keywords in each time period, i.e., the number of papers in each time period that contains a keyword associated with a detected topic. The changes in keyword usage in papers thus serve as an indication of how popular a specific topic or business idea is across different time periods. For this paper, we aim to study the general trend and define each time period as 10 years but this time period can be adjusted accordingly for a different level of analysis.

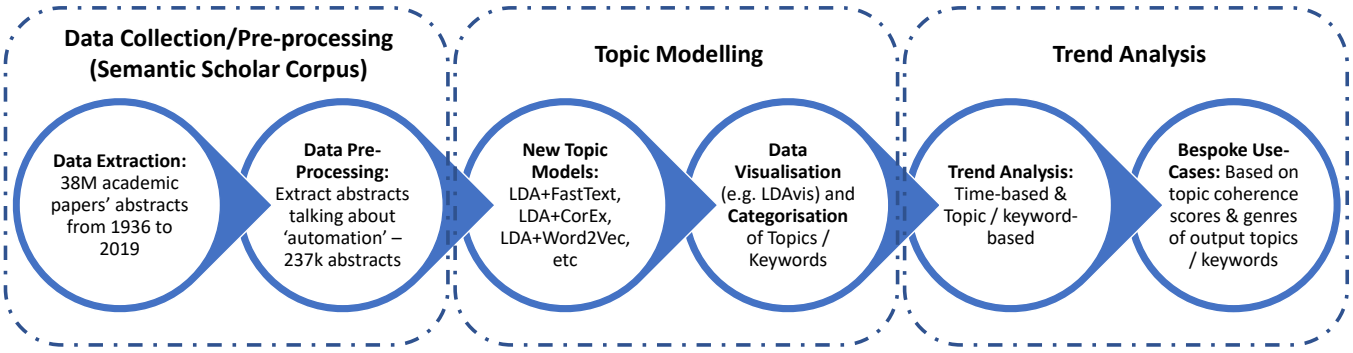

Fig. 1. Overview of framework for identifying business ideas/topics and trends

TABLE I
TOPICS AND KEYWORDS

Topics	Keywords
Unsupervised Machine Learning	data, amplitude, are, explored, extraction, unsupervised, concepts, method, supports, stored
Enzymes	advantages, asynchronous, tuples, enzyme, gives, formally, converts, part, provide, another
Prediction	predicts, key, capable, point, period, computing, independently, appeared, geometry, intersection
Electrical Grids	running, 85, shown, sd, drawback, configuration, substantial, grid, forward, comparatively
Medical Diagnosis	determine, 125, distribution, performs, operates, manner, diagnostic, planar, iterative, ophthalmologists
Sampling	increases, concerning, ref, area, within, accepted, samples, wellknown, thesis, confidential
Supervised Machine Learning	hand, patterns, seeded, supervised, relating, extension, useful, trusted, program, hall
Ototoxic Drugs	code, c, integrated, drug, ototoxic, park, enhanced, ways, give, clarified
Process Efficiency	standardized, triggered, possible, individual, merits, 44, holter, filter, different, enhance
Radio-frequency Sensing/Mgmt.	contribute, functionality, now, steps, chasedb, however, known, radiofrequency, sense, management

III. EXPERIMENTAL SETUP

A. Dataset

To demonstrate the effectiveness of our framework, we performed a case-study using the keyword "Automation" and collected a dataset of 237,258 academic papers to understand the overall trend within the field of "Automation". This keyword was chosen as "Automation" is a prevalent theme across many industries and allows us to better understand the trends in recent years. Our dataset of 237,258 academic papers are those that contain "Automation" in either their title or abstract. We perform the standard data pre-processing steps of keyword tokenization, stopwords and punctuation characters removal on the retrieved abstracts, before performing the remaining steps of our analysis.

B. Algorithms and Baselines

As mentioned in Section II, we proposed various algorithms based on the standard LDA algorithm with a word expansion strategy based on the initial LDA seed keywords with similar words. These similar words are based on a different set of word embedding and expansion algorithms such as CorEx [6], FastText [3] and Word2Vec [4]. This combination of LDA with the various word expansion thus resulted in the following algorithms: LDA+FastText, LDA+CorEx, LDA+Word2Vec, LDA+CorEx+Word2Vec, Doc2Vec+CorEx. In addition to these algorithms, we also compare them to the standard baselines of LDA, CorEx and Non-negative Matrix Factorisation (NMF).

C. Evaluation Metrics

For evaluating how well-clustered each detected topic is, we use the metric of topic coherence which is commonly used in similar works [7]–[9]. Topic coherence is defined as:

$$TC(t, W^{(t)}) = \sum_{w_i \in W^{(t)}} \sum_{w_j \in W^{(t)}, w_i \neq w_j} log \frac{D(w_i, w_j)}{D(w_j)} \quad (1)$$

where $D(w_i, w_j)$ is the number of times both words w_i and w_j appear in the same document, and $D(w_j)$ is the number of times word w_i appear in a document of the corpus.

IV. RESULTS AND DISCUSSION

A. Evaluation based on topic coherence

We evaluate the various algorithms based on minimum, maximum and mean topic coherence score and the results are shown in Table II. We observed that LDA+FastText had the best mean performance at 0.27502, compared to NMF's and LDA's score of 0.2490 and 0.23917, respectively. LDA+FastText also displayed the second highest minimum score at -0.6467, while LDA has the highest minimum score at -0.6464. The remaining algorithms with the other word expansion strategies showed mixed performance. Table I shows

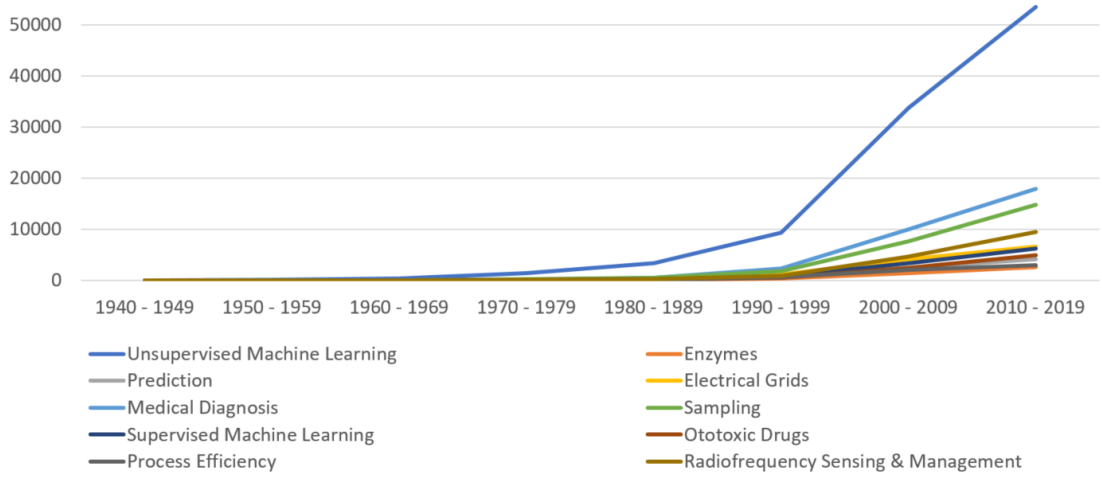

Fig. 2. Trend analysis of topic popularity based on frequency in academic papers.

TABLE II
TOPIC COHERENCE SCORES OF THE VARIOUS ALGORITHMS

Algorithm	Min.	Max.	Mean
LDA+FastText	-0.6467	2.1446	0.2750
LDA+CorEx	-1.2741	2.0516	0.2229
LDA+Word2Vec	-1.5885	2.3107	0.1618
LDA+CorEx+Word2Vec	-1.2514	2.2928	0.1751
Doc2Vec+CorEx	-0.7934	1.5341	0.1812
LDA	-0.6464	2.1452	0.2392
CorEx	-0.9877	0.7835	0.1449
NMF	-0.8768	1.4133	0.2490

an example of some detected topics and their representative keywords.

B. Evaluation based on temporal trends

We also performed a trend analysis on the most popular topics from 1940 to 2019 and found that "Unsupervised Machine Learning" was the most popular topic, while "Medical Diagnosis" is the second most popular within the field of "Automation". This increasing trend started from 1970 for "Unsupervised Machine Learning" and 1990 for "Medical Diagnosis". Figure 2 shows an overview of our trend analysis results, where we observe a gradual increase in popularity for all topics from 1970, particularly for the "Unsupervised Machine Learning" topic.

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed a framework for identifying trending ideas/topics from academic papers, with the key component of a hybrid topic model that combines LDA with various word expansion and embedding methods, namely FastText, Word2Vec, Doc2Vec and CorEx [6]. We demonstrate the utility of this framework on a corpus of academic papers from Semantic Scholar, focusing on the domain of "Automation". There are various future extensions to this work, such as: (i) identifying subject experts from the authors of these academic papers for matching technical experts to relevant businesses/start-ups; and (ii) extending our framework to multiple types of data sources and also mainstream medium like news articles, technical magazines, etc.

ACKNOWLEDGEMENTS

This research is funded in part by the Singapore University of Technology and Design under grant SRG-ISTD-2018-140.

REFERENCES

[1] Sarah Boon, "21st century science overload," Internet, 2017, http://blog.cdnsciencepub.com/21st-century-science-overload/.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[3] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[4] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS'17)*, 2013, pp. 3111–3119.

[5] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31st International conference on machine learning (ICML'14)*, 2014, pp. 1188–1196.

[6] R. J. Gallagher, K. Reing, D. Kale, and G. V. Steeg, "Anchored correlation explanation: Topic modeling with minimal domain knowledge," *Trans. of the Assoc. for Computational Linguistics*, vol. 5, pp. 529–542, 2017.

[7] K. H. Lim, S. Karunasekera, and A. Harwood, "ClusTop: A Clustering-based Topic Modelling Algorithm for Twitter using Word Networks," in *Proceedings of the 2017 IEEE International Conference on Big Data (BigData'17)*, Dec 2017, pp. 2009–2018.

[8] L. Yao, Y. Zhang, B. Wei, H. Qian, and Y. Wang, "Incorporating probabilistic knowledge into topic models," in *Proceedings of the 19th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'15)*, 2015, pp. 586–597.

[9] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie, "Improving LDA topic models for microblogs via tweet pooling and automatic labeling," in *Proceedings of the 36th Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'13)*, 2013, pp. 889–892.