Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 177 (2020) 581–586

Procedia Computer Science

www.elsevier.com/locate/procedia

International Workshop on Artificial Intelligence for Natural Language Processing
(IA&NLP 2020)
November 2-5, 2020, Madeira, Portugal

A Framework for Understanding the Dynamics of Science: A Case Study on AI

Sahand Vahidnia[a,*], Alireza Abbasi[a], Hussein A. Abbass[a]

[a]*School of Engineering and IT, UNSW, Canberra ACT 2612, Australia*

Abstract

The Science of Science is an emerging field that enables tracking the dynamics of science in the form of birth and death of scientific fields and research trends in them. This study proposes a framework to analyze the content of academic publications to extract research trends and explore their temporal evolution. In this proposed framework, state-of-the-art embedding techniques are reviewed and utilized to consider semantic vectors of publications' keywords, which have been used to produce semantic-based clusters reflecting research trends or sub-fields. We compare our proposed method with LDA, as a baseline method, to demonstrate the explainability of the clusters, applying them to the field of "Artificial Intelligence" (AI) as a case study.

Keywords: Dynamics of Science; Science Mapping; Word Embedding; Artificial Intelligence

1. Introduction and Background

Detecting and predicting scientific trends which consist of the birth, growth, and decline of scientific fields are very important in planning future research projects and studies [17]. There have been varying methods proposed and explored in the literature to analyze and understand the dynamics of science considering the change of scientific fields and their sub-fields. Topic modelling techniques such as Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA) are amongst the most popular methods in the field that are used to understand relationships among data and text documents [5]. Topic modeling techniques are used to uncover the conceptual *topics*, as a group of words, from set of texts considering various statistical analysis of words in texts. Although these methods can successfully cluster words, the clusters of words may not be interpretable by humans. These clusters might be semantically and epistemologically irrelevant to each other as it is observed in this study. This problem can be much more severe in the

[*] Corresponding author. Tel.: +61-497633317 ;
E-mail address: s.vahidnia@unsw.edu.au

10.1016/j.procs.2020.10.082

case of having short documents and texts [9]. After the recent developments in NLP, and development of embedding techniques and classification methods like word2vec [10], Glove [12] and FastText [6], this has been more possible than ever to automatically classify the fields of science and analyze its dynamics.

Many previous studies in the field rely on word co-occurrences to map the scientific fronts. A relatively early and a very influential study in co-word networks have been conducted by [1] analyzing research topics based on co-occurrences of word-reference combinations. [2] utilizes co-word analysis to reveal the structure and development of research fields. For this purpose, factor analysis, cluster analysis, multivariate analysis and social network analysis, using the matrix of word co-occurrences have been implemented using project funding data. K-core analysis has been used in the study to detect the themes in the field. [20] performed three different co-word network analysis, using WoS KeywordPlus and Pajek software. There also have been other studies utilizing similar techniques in other fields like [16], finding research trends about vitamin D.

In addition to word co-occurrence approach, others utilized words and context for science mapping and topic front clustering. Zhang et al. [18] proposed a topic-based model, utilizing LDA and scientific evolutionary pathway modeling (SEP). The study uses LDA to profile the articles published in Knowledge-based Systems journal, generating 25 topics for the 2566 articles. Later, the relationship among these topics have been evaluated using co-topic networks. SEP has been used for identifying and analyzing topics and their relationships (formerly studied in [19]), in a sequential time period. SEP follows a similar approach to a typical clustering algorithm, but in a sequential and temporal order. It starts with an initialized topic and assigns new articles to topics using Salton's cosine measure, then updates the clusters with streaming data. In a recent study by Park et al. [11], a similar method proposed that uses cosine similarity to cluster the documents. Unlike SEP, Park et al. utilized more advanced document embedding than TF-IDF and other similar frequency related features, to capture the semantics. In addition, a deep document clustering approach has been taken to overcome the deficiencies seen in previous methods.

In this study, we propose a framework (FTSci) to utilize the strength of word embedding techniques, to include semantics into scientific topic front clustering and field dynamic analysis. To track the temporal evolution a field, the similarity among the clusters over different time periods are calculated and visualized. As a case study, we use our developed framework to analyse the dynamics of "Artificial Intelligence" publications to extract its sub-fields and their dynamics over time.

2. Methodology

The methodology comprises two main sections: clustering, and mapping. At the initial section, a framework is proposed to cluster author keywords as knowledge entities, to extract research trends by historical data. The second section will explain the procedure in the framework to monitor the temporal evolution of research trends.

Most studies in the field have limited their data to specific journals or fields while interdisciplinary fields often occur among various journals from different fields.For example, excluding psychology from social network analysis will limit findings in the field due to the connections between the two fields. To overcome this shortcoming, we use a different method of data gathering scheme using a keyword as a search query. In addition, the data in this study is gathered for a relatively long period to properly capture the evolution of a field. The academic publication data in this study has been gathered from two different indexing services: Scopus and Web of Science (WoS), with "artificial intelligence" as the search term in titles, abstracts and keywords. The search yielded around 310k records in Scopus and 37k records in WoS. It was noticed that WoS records are more consistent regarding their meta-data, including the availability and organization of author keywords. Therefore the records (author keywords) from WoS is used as the main dataset for the case study. The data then is divided into six periods, enabling us to study the evolution: 1990-2004, 2005-2007, 2008-2010, 2011-2013, 2014-2016, and 2017-2018. Scopus data is merged into WoS data, yielding a new and larger dataset, which is used solely for training purpose as a large corpus. For training purpose, abstracts and titles from the merged dataset has been extracted, divided to sentences, pre-processed and recorded in a dataset of sentences. The text pre-processing procedure comprises lemmatization, removal of symbols, word stripping, word expansion and synonym elimination by a custom thesaurus file.

2.1. Research Trend Clustering

There have already been studies to automatically categorize or group research trends [16] [2] [1]. Yet they mostly rely on statistical methods and/or network attributes of entities such as co-word or citation networks. In this study, we are proposing a framework to leverage the strength of word embedding techniques when categorizing words. In this section, we define the research trends by their corresponding keywords. For this purpose, only author provided keywords will be utilized to capture the mindset of authors for research-front clustering.

Word Embeddings: Needless to say, the main ingredient of text clustering techniques is to represent the data in vector space, by obtaining word embeddings. Word embedding techniques benefit from neural networks to generate embedding vector representations of words [7]. Regarding the clustering task, it has been demonstrated in prior studies that neural embeddings outperforms other techniques [3]. Hence, in this study, to acquire word embeddings, FastText [6] model is used to obtain word embeddings. FastText is very similar to word2vec [10] in nature, with a few more capabilities. The FastText model leverages a single layer neural network, which turns it into a linear model. On the other hand, this also makes it very fast and simple. A feature of FastText which makes it stand out in comparison to similar methods is the utilization of sub-word features and n-grams. Character level features have been explored in more complex methods such as BERT, but they usually lack the speed, efficiency and accuracy of FastText. Another advantage of using FastText instead of BERT is having static word vectors, regardless of the context of that word. This enables us to obtain isolated word embeddings, as the author keyword data lacks context.

To train and use FastText, a fake classification training task is created to train the neural network weights and save them as word embeddings. The small size of the neural networks in FastText, makes it even easier to configure and customize dimensionality of the embeddings, which puts FastText in a further advantageous position comparing to BERT and larger models.

Word Vectors: FastText as a language model, originally has been trained on Wikipedia or Common Crawl data, generating 300 dimensional word representation models. In this study, the scope and the context of the goal keywords are far more limited than a general language problem. Thus, having 300 dimensions is very likely to reduce the clustering performance [14] and confronting the curse of dimensionality is very likely. In this case, traditional similarity measures may not perform well [8]. Hence, FastText model is trained on aforementioned merged sentence data using Gensim library [13], generating 15 dimensional representations. For the clustering task in this study, the aim is clustering based on scientific representation of terms. However, not all features (dimensions) are helpful in this task. Therefore, a feature elimination task can improve the clustering performance. Lower number of dimensions is preferred in this study, as the vocabulary size is relatively small and higher dimensions will add unnecessary amount of complexity to the clustering task. Feature selection and elimination is carried out by eliminating the features with very low variance and also removing the dimensions with high variance in a meaningless or wrong dimensions (e.g. a dimension which indicates the grammatical aspect instead of meaning). After this feature selection, 10 features are nominated and kept for clustering task. It is worth mentioning that the word vectors produced in embedding techniques tend to be close to each other and sometimes it will be difficult to work with them.

Clustering: Prior to clustering, to reduce the amount of noise in data, here keywords with low frequencies are eliminated. Thus, top three percent of the keywords are selected for clustering, which yields keyword frequency thresholds around 8 to 10 repetitions, that leads to keeping only about 2000 keywords for the data of years 1990 to 2018. In the clustering task, hierarchical agglomerative clustering with Ward's method [15] is used. This is a bottom-up hierarchical clustering technique, which minimizes the total within-cluster variance. Hierarchical clustering can provide number of benefits and flexibilities like automatically deciding on number of clusters. However, the automatic number of clusters might not always be optimal for our task. Number of clusters are selected with supervision and observation of the dendrograms, which facilitate the cluster number determination task by visualizing the distances of each cluster in a hierarchical representation (See Figure 1.). For example by drawing a horizontal line on 400 mark in y-axis (reflecting the distance of clusters), it will cut four vertical lines in the tree, meaning we can get 4 clusters. If we go lower in the hierarchy (e.g. distance 300) we can get 7 clusters.

2.2. Research Evolution Mapping

In order to visualize temporal evolution of research trends, we track the extracted clusters across time periods. For this purpose, clustering is performed on all six selected periods, in a cumulative fashion. Although being cumulative,

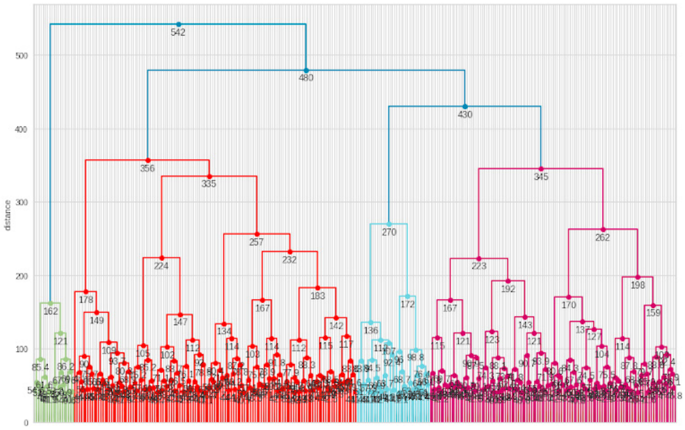

Fig. 1. Truncated dendrogram sample for hierarchical clustering

each period is mostly affected by its new content, as the new content is large in size and dominates the overall context. In addition, vocabulary and content growth rate for all the periods after 2010 is non-linear and increasing.

To track the clusters, centroid of each cluster is calculated as element wise mean of the word embeddings in the cluster. Then the distance of the centroids are measured to find similarity of each cluster in a time period (t) with all clusters in the following time periods ($t + 1$) and ($t + 2$). This allows us to determine which clusters are the children of the previous clusters.Measuring this distance for ($t + 2$) enables us to detect the resurrecting trends. This measurement has been done in form of cosine similarity measurement ($1 - cosine\_distance$). This has been chosen over Euclidean distance for better normalized representation. As it was mentioned earlier, due to closeness and low distance of the word vectors from each other, there are a degree of similarity among almost all pairs of clusters. To overcome this issue, we first measured the intra-period cluster center distances, using 90% of this distance as a threshold for inter-period similarity measurements. This enables us to visualize the flow of the topics as links among clusters. Labeling is the final task in the framework. For labeling of clusters, a standard, yet very simple procedure has been adapted, where top 98 percentile of the cluster keywords (sorted by frequency) are used as labels of the clusters.

3. Results

Research Trend Clustering: In order to validate our proposed method, we are comparing our results to LDA as a baseline topic modeling method. Similar to the work by [18], and to get the optimal results from LDA, n-grams are connected by an underscore to treat them as single words. As illustrated at Table 3, our proposed methods can generate topics in the scientific fields which are linguistically more relevant to each other, compared to the topics generated by LDA. In addition, crisp clustering prevents the repetition of keywords in clusters, whereas LDA topics suffers gravely from this due to its bag-of-words nature. For example, “artificial neural networks” is present in most topics of LDA and makes the topics very difficult to distinguish. FTSci generates more distinguishable and relevant clusters, like putting “power” and “energy” in the same cluster, or “Decision making” and “Decision support system” together. Tagging and labeling the clusters in FTSci is comparatively much easier, and semantically coherent, as similarly stated by prior studies [4].

Research Evolution Mapping: The real strength of the embedding-based clustering in this study is leveraged at temporal evolution analysis. Having good embeddings makes it possible to better reveal and represent semantic structures. As illustrated in Figure 2, links among clusters are meaningful and explainable. In this illustration, link strengths among clusters are proportional to link widths and their opacity values. Using this diagram, one can interpret the temporal evolution, like the links among “genetic algorithm” clusters in all periods and their inter-connections with “optimization” and “machine learning”. Some insights from the temporal evolution of research trends in AI includes:

- “Decision Support System” was a major field until 2010 and morphs into the fields around “Robotic”, “Scheduling”, “Planning” and “Expert System”.

Table 1. Comparison of clusters: LDA vs. Our proposed framework (FTsci) for WoS author keyword clustering from 1990 to 2018.

FTSci	LDA
* ML - fuzzy logic - data mining	ANN - GA - SVM
* planning - control - scheduling	ML - ANNs - DL
* robotics - agent - internet thing	reasoning - GA - ANNs
* SVM - simulation - multi agent	CNN - DL - distributed AI
* ontology - knowledge - HCI	IOT - robotics - multi agent
* decision making - multi agent system - DSS	reasoning - distributed AI - prediction
* GA - optimization - algorithm	ontology - scheduling - semantic web
* ANN - distributed AI	data mining - ML - expert
* expert system - knowledge system - intelligent agent	ANNs - expert system - fuzzy logic - ML
* expert - reasoning - knowledge representation	optimization - GA - simulation
* power - FPGA - energy	ANNs - fuzzy logic - ML
	ANNs - GA - ML
	emotion - PCA - SVM
	ML - NLP - multi agent system

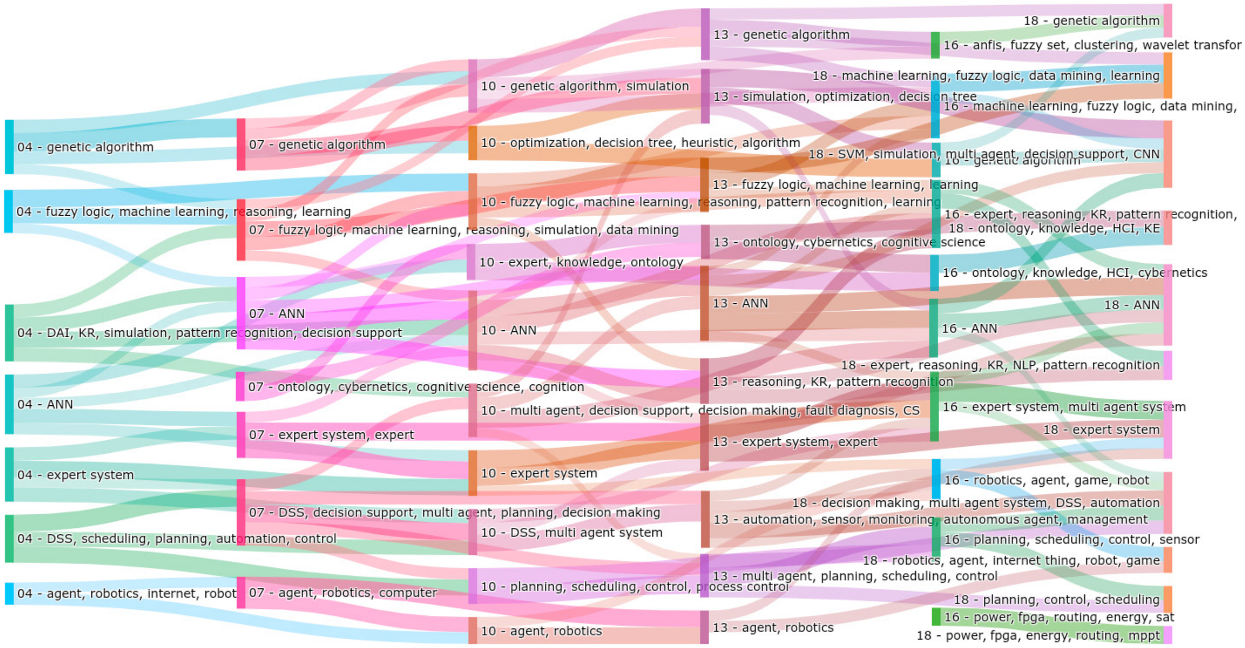

Fig. 2. Sankey diagram, illustrating the temporal evolution of research trends over time, starting from 1990-2004 (left) to 1990-2018 (right). Initial digits for each label are indicative of the year period. clusters in the same period have similar colour, but cross-period colors are random.

- “Fuzzy Logic” appears very close to “Machine Learning” in all periods and it stands out until 2013, at top of the cluster. At around 2016 “Machine Learning” overtakes ‘fuzzy logic” and takes the top spot in clusters.
- “Power”, “routing” and “Energy” starts to appear in AI between 2016 until 2018. Perhaps due to more focus on the applications of AI in Energy domain.
- “Optimization” in AI starts to stand out around 2010, although existed in prior periods, with roots in fields around “Genetic Algorithm” until 2013 and merges into “Artificial Neural Networks”, “Support Vector Machine” and “Genetic Algorithm” again.
- “Automation” and “Sensor” became important since 2013, but morphs into “Scheduling”, “Planning” and “Expert System” right after. It only lives for a single period as top of the fields around.

4. Discussion and Conclusion

In this study, we proposed a framework to detect research trends over time and visualize temporal evolution of AI as a sample field of study. The framework uses word vectors, acquired by fastText model, for clustering which are used to visualize and analyze temporal dynamics of science. In our experiment, we trained our model on relevant publications' abstracts and titles providing better embeddings in comparison to the pre-trained model on millions of data point. It was also observed that lower number of dimensions are far easier to manage. In 15 dimensional data, revealing and eliminating the misleading and unnecessary features was manageable. In higher dimensions, due to the curse of dimensionality, nodes get pulled apart and clustering performance decreases. To remedy this issue, the number of samples can be increased, which might also increase the noise levels. Yet, due to lack of feature elimination, cluster quality is not on par with lower dimensions. In this study, we show that smart embeddings, such as FastText word vectors, can be very beneficial in revealing research trends, and generally exposing semantic structures. The proposed framework and procedure in this study reached its target in our experiment. Regardless of the relatively small training data, the framework comprising the steps of clustering the author keywords and extraction of research trends of each time period, and linking the research trends across periods successfully executed.

The possibility of a smart and automatic feature selection can be discussed in future works, as it was observed in our experiments that the most misleading feature was the principal component of the vector. However, this requires further investigation and validation.

References

[1] Van den Besselaar, P., Heimeriks, G.: Mapping research topics using word-reference co-occurrences: A method and an exploratory case study. Scientometrics **68**(3), 377–393 (2006)

[2] Chen, X., Chen, J., Wu, D., Xie, Y., Li, J.: Mapping the research trends by co-word analysis based on keywords from funded project. Procedia Computer Science **91**, 547–555 (2016)

[3] Curiskis, S.A., Drake, B., Osborn, T.R., Kennedy, P.J.: An evaluation of document clustering and topic modelling in two online social networks: Twitter and reddit. Information Processing & Management **57**(2), 102034 (2020)

[4] Hu, Z., Luo, G., Sachan, M., Xing, E.P., Nie, Z.: Grounding topic models with knowledge bases. In: IJCAI. vol. 16, pp. 1578–1584 (2016)

[5] Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., Zhao, L.: Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. Multimedia Tools and Applications **78**(11), 15169–15211 (2019)

[6] Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759 (2016)

[7] Kim, J., Yoon, J., Park, E., Choi, S.: Patent document clustering with deep embeddings. Scientometrics pp. 1–15 (2020)

[8] Kriegel, H.P., Kröger, P., Zimek, A.: Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. ACM Transactions on Knowledge Discovery from Data (TKDD) **3**(1), 1 (2009)

[9] Li, C., Wang, H., Zhang, Z., Sun, A., Ma, Z.: Topic modeling for short texts with auxiliary word embeddings. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. pp. 165–174. ACM (2016)

[10] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)

[11] Park, J., Park, C., Kim, J., Cho, M., Park, S.: Adc: Advanced document clustering using contextualized representations. Expert Systems with Applications **137**, 157–166 (2019)

[12] Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)

[13] Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50. ELRA, Valletta, Malta (May 2010)

[14] Steinbach, M., Ertöz, L., Kumar, V.: The challenges of clustering high dimensional data. In: New directions in statistical physics, pp. 273–309. Springer (2004)

[15] Ward Jr, J.H.: Hierarchical grouping to optimize an objective function. Journal of the American statistical association **58**(301), 236–244 (1963)

[16] Yang, A., Lv, Q., Chen, F., Wang, D., Liu, Y., Shi, W.: Identification of recent trends in research on vitamin d: A quantitative and co-word analysis. Medical science monitor: international medical journal of experimental and clinical research **25**, 643 (2019)

[17] Zeng, A., Shen, Z., Zhou, J., Wu, J., Fan, Y., Wang, Y., Stanley, H.E.: The science of science: From the perspective of complex systems. Physics Reports **714-715**, 1–73 (2017). https://doi.org/10.1016/j.physrep.2017.10.001

[18] Zhang, Y., Chen, H., Lu, J., Zhang, G.: Detecting and predicting the topic change of Knowledge-based Systems: A topic-based bibliometric analysis from 1991 to 2016. Knowledge-Based Systems **133**, 255–268 (2017). https://doi.org/10.1016/j.knosys.2017.07.011

[19] Zhang, Y., Zhang, G., Zhu, D., Lu, J.: Scientific evolutionary pathways: Identifying and visualizing relationships for scientific topics. Journal of the Association for Information Science and Technology **68**(8), 1925–1939 (aug 2017). https://doi.org/10.1002/asi.23814

[20] Zhao, W., Mao, J., Lu, K.: Ranking themes on co-word networks: Exploring the relationships among different metrics. Information Processing & Management **54**(2), 203–218 (2018)