Document Clustering and Labeling for Research Trend Extraction and Evolution Mapping

Sahand Vahidnia
s.vahidnia@unsw.edu.au
School of Engineering and IT, UNSW
Canberra, Australia

Alireza Abbasi
a.abbasi@unsw.edu.au
School of Engineering and IT, UNSW
Canberra, Australia

Hussein A. Abbass
h.abbass@unsw.edu.au
School of Engineering and IT, UNSW
Canberra, Australia

ABSTRACT

In this study, a method is being proposed to extract research trends and their temporal evolution, throughout discrete time periods. For this purpose, a document embedding method is developed, adapting contextualized word embedding techniques. The method utilizes published academic documents as knowledge units, then clusters them into groups, each representing a series of related fields of research. Various labeling techniques are also explored, including source title popularity, author keyword popularity, term popularity, term importance, and Wikipedia-based automated labeling to evaluate the quality of clusters and explore their explainability. A case study is conducted on Artificial Intelligence (AI) related publications, putting the method to test and observe the evolution of AI within the studied periods. In this study, we show that utilization of neural embeddings in conjunction with paragraph-term weights would provide simple, yet reliable paragraph embeddings, that can be used for clustering of the textual data. Additionally, we show that cluster centroids can be used for cluster tagging, labeling, and inter-connecting for topic evolution study.

CCS CONCEPTS

• **Information systems** → **Data mining**; **Document topic models**; • **Computing methodologies** → *Topic modeling*.

KEYWORDS

Dynamics of Science, Science Mapping, Text Embedding, Artificial Intelligence

1 INTRODUCTION

Understanding and predicting future discoveries and scientific achievements is an emerging field of research, which involves scientists, businesses, and even governments. This field is also known as Science of Science (SciSci), which aims to understand, quantify and predict scientific research dynamics and the drivers of that dynamics in different forms such as the birth and death of scientific fields and their sub-fields [1] that can be identified by tracking the changes of research trends. A field / sub-field may go through different stages, which consist of the birth, growth, and decline of scientific trends. The initial stage or the birth of a sub-field may come from splitting and merging of other fields. Later, a field may attract more researchers and observe growth or can decline, as sociologists believe scientists either take a risky approach to make novel research or they prefer to stay on the safe side and stick to tradition [2][3]. 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 modeling techniques such as Latent Dirichlet Allocation (LDA) [4] 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], and network analysis techniques such as co-occurrences of words, citation networks are one of the most explored methods in the literature for revealing relationships in data. However, after recent developments in machine learning and natural language processing (NLP), new methods in text mining such as word and document embeddings have facilitated analyzing the metadata or contents of publications in different fields to understand the dynamics of those fields.

Understanding the dynamics of science and the ability to predict these dynamics and evolution of a field of science, helps us to understand if there is something important left behind accidentally or if there is a branch of science at a phase transition moving towards a major discovery. The ultimate objective of this research is to deepen our understanding of the dynamics of science and develop methods and frameworks to make the historic analysis of science dynamics and temporal evolution possible and automated, and making predictions for future evolution possible for the scientific community. The objectives of this research can be divided into the following two main categories:

The objective of this study is to detect and map scientific trends. Revealing these trends requires us to exploit contextual features in the scientific research domain and understand its dynamics. In this study we propose a simple framework to facilitate the exploration of scientific trends and their evolution, utilizing contextual features and deep neural embeddings. Our proposed framework is then applied in a case study to understand the path of scientific evolution in artificial intelligence. In this study, we show how the trends and topics in science can be extracted using document vectors and extraction of context. The overall outline of the proposed framework is as illustrated in Fig.1.

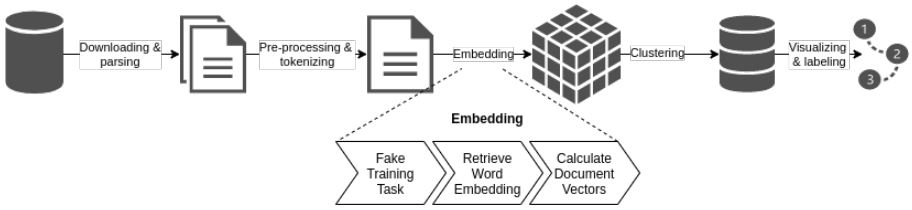

Figure 1: Flowchart of the proposed framework outline.

2 LITERATURE REVIEW

Many previous studies in the field rely on word co-occurrences to map the scientific fronts. A relatively early and very influential study in co-word networks has been conducted by Van Den Besselaar et. al [6] analyzing research topics based on co-occurrences of word-reference combinations. They put the structure of science into four levels: (1) discipline (e.g., computer science); (2) research field (e.g. AI); (3) sub-field (e.g., machine learning); and (4) research topics (e.g., deep learning). Sedighi [7] has analyzed the research areas, their relationships, and growth trends in the field of Informetrics using word co-occurrence. Chen et al. [8] in a study utilize 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 performed. It used the meta-data of 2054 funded projects from 2011 to 2015 and only the keywords having more than 8 repetitions are considered (6,153 keywords). Authors have used Matlab to get the co-occurrence matrix and other similarity analyses, including the co-correlation matrix, have been done using UCINET and further SNA have been done using VOSviewer. Zhao et al. in a study [9] seek to find the relationships among different theme ranking metrics comprised of frequency-based and network-based methods. The study categorizes the metrics into three groups: (1) degree centrality, H-index, and coreness, (2) betweenness centrality, clustering coefficient, and frequency, and (3) weighted PageRank. The study suggests that recently co-word analysis has shifted to network-based metrics and attempts to examine the relationships among these metrics of term ranking. In the empirical phase of the study, Keywords Plus from WoS data has been used instead of extracting keywords from the text and using author keywords, as many author keywords are missing in data. These keywords have been used in the co-word analysis in the aforementioned three fields, using the Pajek tool. There also have been other studies utilizing similar techniques in other fields like Yang et al. [10], which is a study of finding research trends about vitamin D.

In a study of knowledge evolution detection and prediction, Zhang et al. [11] propose a topic-based model, utilizing LDA and scientific evolutionary pathway modeling (SEP). The study uses LDA to profile the articles published in the Knowledge-based Systems journal and generates 25 topics for the 2566 articles. An interesting workaround is suggested in this study, which is to concatenate the n-grams to form a uni-gram which bypasses the preference of LDA in single words. This workaround has also been used in other methods including word2vec [12]. Later, the relationships 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 [13]), in a sequential time period in this study. SEP follows a similar path to a typical clustering algorithm, but in sequential temporal order. The study utilizes Salton's cosine measure [14] to assign topics and articles. The study acknowledges that using word embedding techniques could improve the result, instead of term frequency based vectors. In another study [15], they continue the previous work [11], using Word2Vec [12] as embedding technique, coupled with a kernel k-mean clustering algorithm. As have been shown in many other studies, word2vec and other embedding and language models can exploit more complex features in textual data. Hence, it can exploit the desired features for clustering purposes. In the experiments conducted in this work, pre-trained 100-dimensional vectors have been utilized. As for clustering, a polynomial kernel k-means with cosine distance measure have been adapted to better cluster bibliometric features.

In a study of technology trend monitoring [16], a framework is suggested to use patent data in conjunction with Twitter data. Due to the lag in the patent data and not capturing the whole technological advances, utilization of Twitter data is being suggested in this work, which comprises many technological discussions, prior to their publication. The clustering in this study has been carried out using Lingo algorithm. The study uses Carrot2 workbench for visualization of patent clusters. Then author-topic over time (ATOT) model is used to analyze the tweets and obtain topic-feature words probability distribution and topic-user probability distribution. Finally, in a recent review study [17], different document clustering and topic model methods are compared and evaluated. The study confirms the advantage of advanced embedding methods in contrast to traditional methods like *tf-idf*. The study claims that methods like doc2vec [18] with *tf-idf* weights would outperform other methods. They also show that it is possible to readily use doc2vec in with k-means clustering.

3 METHODOLOGY

3.1 Data Collection

The language model training data has been collected via Scopus from 1990 to 2019, by "artificial intelligence" query search key in titles, abstracts, and keywords fields, yielding 310k records (dataset A). This collection method allows us to increase the variance in training data, resulting in better generalization. In contrary, the main data for the analysis has been collected from three mainstream journals in AI from 1970 to 2019 (dataset B): "Artificial Intelligence" (2575 records), "Artificial Intelligence Review" (890 records), and

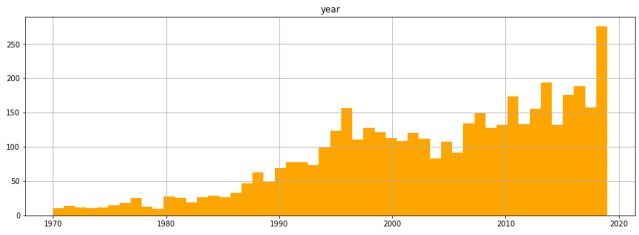

Figure 2: Data growth from 1970 to 2019 in the three journals, yielding over about 4300 records.

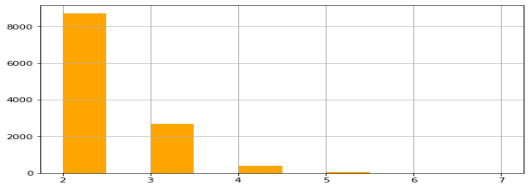

Figure 3: N-gram dictionary histogram of the dataset. Y-axis is frequency and X-axis is N.

"Journal of Artificial Intelligence Research" (1006 records). The reason for excluding other journals is to limit or eliminate the bias in some journals towards specific applications (e.g. health) or approaches (e.g., engineering and deep learning) in AI. As observed in the Web of Knowledge master journal list categories, these three journals were selected to best fit the purpose of this research, having the minimum bias to specific applications or approaches. Fig. 2 illustrates the data records throughout the period of the study data.

3.2 Data Pre-processing

After the acquisition of data, the following initial pre-processing steps were conducted on the datasets A and B: (1) Removal of duplicated records by their Digital Object Identifiers (DOI). (2) Removal of records with missing abstracts. (3) Concatenating the titles and the abstracts: using the title as the initial sentence of an abstract, then saving them in the abstract column. (4) Lemmatization of abstracts: Noun level lemmatization and skipping the other parts of speech. (5) Replacement of very famous acronyms with their corresponding terms. (6) Removal of words like 'et al.', 'eg.', 'ie.', and 'fig.', which generally have tailing dots and would hamper with the sentence extraction process, without carrying meaningful information. (7) Converting all British English words to American English words for consistency of the data. (8) Removal of punctuation, special characters, and numbers. (9) Sentence extraction for training the language model (dataset A only).

The secondary pre-processing stage is as follows, which is carried out on the analysis data (dataset B) only. This data is only used in the labeling stage, which will be denoted "label data": (1) Removal of stop words, like "a" and "the" from the corpus. (2) Concatenation of n_grams based on the taxonomy generated from author keywords in the dataset A, by replacing spaces with underscores (artificial intelligence -> artificial_intelligence). This taxonomy only contains the 95 percentile of keywords n-gram keywords, to cover the most important keywords. Hence, n-gram keywords with a frequency of at least six are kept and the rest are ignored. In addition, a condition of M>2N has been maintained to keep the keywords with N-grams and M characters. This eliminates the keywords with characters counts lower than 2N, which usually are generic words and potentially harmful for the data and the text corpus. For this purpose, N-grams are sorted from higher N to lower N, then replaced in the corpus with corresponding words. In this study, $N \in \{1, ..., 6\}$, as numbers over 6 are usually either errors or very sparse (please refer Fig. 3 for the histogram of N in n-grams). To eliminate any chance of mid-word overwriting, all searches are done by leading and ending spaces, and to make the corpus suitable for this, a leading and ending space is added to all data records. (3) Data is divided into nine periods by publication year, to [1970,1989], [1990,1994], [1995,1999], [2000,2004], [2005,2007], [2008,2010], [2011,2013], [2014,2016], and [2017,2019] periods. This division into nine periods provides us with more uniform count of records within each period, facilitating the clustering approach.

3.3 Contextualized Embedding for Document Clustering

Frequency-based analyses are not the only ways to cluster documents for understanding their topics. Another way to get the topics within a set of documents is to use contextualized embeddings. As the name suggests, this provides further context awareness to the approaches of uncovering topics and latent information in the text. There also have already been studies to automatically categorize or group research trends [10] [8] [6]. Yet they rely on statistical methods and/or network attributes of entities such as co-word or citation networks. In this study, we are leveraging the strength of contextualized embedding techniques when categorizing documents. Later, we define the research trends by their corresponding keywords. To facilitate this in labeling, authors' keywords is utilized to enhance the context and capture the mindset of authors for research-front clustering.

3.3.1 Embedding Method. Needless to say, the main ingredient of text clustering techniques is to represent the data in vector space. Word embeddings or vectors have been around for a long time. Simple word vectors are one-hot embeddings like bag-of-words. However, they don't provide much information regarding the data. Many methods incorporate simple statistical vectors like *tf-idf* [19] or bag-of-words. That is why models like Word2Vec (W2V) [12] were introduced. Regarding the clustering task, it has been demonstrated in prior studies that neural embeddings outperform other embedding techniques [17]. W2V is a single hidden layer neural network and works with two different models: Common Bag of Words (CBOW), and skip-gram model. CBOW model tries to predict a word based on its context (surrounding words). Word embedding techniques benefit from neural networks to generate embedding vector representations of words [20]. The method of choice for generating vectors in this study is FastText [21], due to its richer embeddings. FastText is very similar to word2vec in nature, with few more tricks. FastText also leverages a single layer neural network, which 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 too, but they usually lack the speed, efficiency, and accuracy of FastText. Yet, there exist models that can outperform FastText, such as BERT [22], preserving features from bi-directional word orders and sub-word information. BERT is far more complex and resource-intensive than FastText in training and fine-tuning, providing vectors of very large dimensions.

3.3.2 Word Vectors and Dimensions. It is preferred to utilize low dimension size in embeddings as opposed to the original dimension size of pretrained FastText embeddings. The curse of dimensionality is known to be a common problem when dealing with similar tasks [23] [24]. Gensim library [25] has been used here to generating 50-dimensional FastText models, using the large corpus (dataset A). It has been concluded empirically that 50 would be an optimal dimension size, dealing with clustering tasks of this scale. The word embedding task is a fake training task, to retrieve word weight from the neural networks, which is used as the word embeddings. Thus, each dimension may represent a specific feature of a document or text. Due to the complexity of the dimensions and their meanings when dealing with document and text clustering, no manual feature engineering is carried out. Additionally, no further dimensionality reduction is used, as it is possible to select the output size of the neural network in FastText, rendering further utilization of auto-encoders and similar methods less useful. Simpler dimensionality reduction methods like PCA were also attempted, yielding in sub-optimal results. It was observed that dimensionality reduction techniques for this task have little to no positive effect. Hence the raw FastText embeddings are preferred in this study.

3.3.3 Document Embedding and Vectors. As we aim to cluster documents based on their scientific representation, author keywords are ignored and the embedding is based solely on document titles and abstracts. For this task, document vectors are required to be calculated. There have been a number of studies to calculate document vectors and document clustering, including [26] [27] [28]. An intuitive method is to average the word representations to acquire document vectors. However, it won't provide stable results. Arora et al [28] provide a baseline method, which we have adopted in this study for document embedding. The method is called "Smooth Inverse Frequency" (SIF). SIF is basically a weighted averaging method, based on probability and inverse frequency for words in documents and is claimed to have 5 to 13% improvements, thus is adapted in this study. The SIF adaptation in this study is illustrated at the following equation, where $wv(t)$ is calculated for each term for all strings, and then is divided by the number of terms in the corresponding string. Here $v(t)$ represents each term vector, and $p(t)$ is the probability of seeing that term. Regarding the α, the constant value of $1e-3$ is used.

$$w(t) = \frac{\alpha}{\alpha + p(t)} \tag{1}$$

$$wv(t) = w(t) * v(t) \tag{2}$$

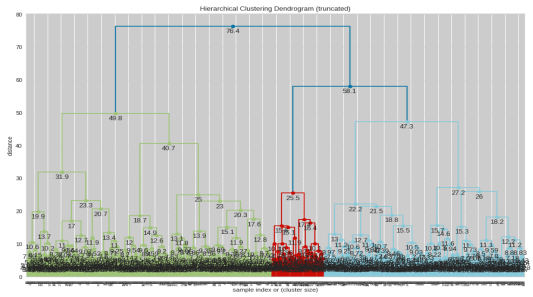

Figure 4: Dendrogram of document vector clusters in the period of 1970 to 2019.

3.3.4 Clustering Approach. Comparing to other well-known clustering methods like k-means, it was observed that the most successful clustering technique to fit our data and task is hierarchical agglomerative clustering with Ward's method [29]. This is a bottom-up hierarchical clustering technique, which minimizes the total within-cluster variance. Hierarchical clustering can provide a number of benefits and flexibilities like decision support on the number of clusters. The selection of cluster numbers for hierarchical clustering usually can be done via a dendrogram. Dendrogram basically shows the hierarchical structure of the nodes, based on their closeness to each other, as illustrated in Fig. 4.

3.3.5 Cluster Labeling Approach. Cluster labeling has been carried out using two different methods. The initial method uses important words within the text. Using this method, clusters are tagged using a normalized *tf-idf* scoring method to extract the important words within each cluster by providing further discrimination to cluster term content, from count vectorization of terms with less than 0.8 presence in documents. The following equation shows the scoring technique for cluster tags.

$$score(t, c) = tf(t, c) * icf(t) \tag{3}$$

$$icf(t) = log\frac{(1+n)}{(1+cf(t))} + 1 \tag{4}$$

Where t is a term, c is the corresponding cluster, cf is the frequency of clusters with term t, and n is the number of clusters.

These top term tags can help us identify the topic and subject area of each cluster and its cover. These tags are used to extract the important terms within each cluster and can be used to roughly estimate the overall cluster label. In other words, these terms summarize the context of each cluster in a couple of keywords. However, this can only loosely define the labels and fields, without any formal definition, which renders it less useful, unless used in conjunction with expert opinion. To address this problem, another method is developed to label clusters, utilizing the definitions within "Outline of artificial intelligence" in Wikipedia. Hence, all pages from both "Applications" and "Approaches" of AI in this outline are parsed and vectorized, yielding a single 50-dimensional vector for each "Application" and each "Approach". The vectorization steps are as follows: First, each page is parsed, all unnecessary data noise, including references and titles, are cleaned. Then each page is turned into a corpus of sentences and each sentence is embedded individually

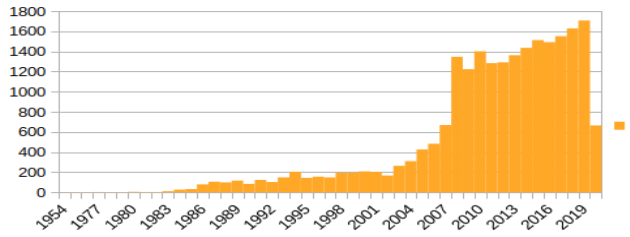

Figure 5: Query result for studies of decision support system and AI trend. Y-axis is Frequency.

using the SIF method and parameters, introduced at 3.3.3. Finally, to obtain the page vector, the mean of all sentence vectors are calculated, providing us with the centroid of the Wikipedia document, or in other words, the location of approaches and applications of AI in our vector space. To utilize these vectors for obtaining Wikipedia approaches and applications of AI, the centroid of each cluster is compared to each of the approaches and applications of AI, using their cosine similarities, as defined in equation 5. Cosine similarity has been used for NLP tasks [30] [11] as it is known to be the best measure fitting this task. The top two closest approaches and applications are then selected as labels for each cluster.

$$sim(a, b) = \frac{a.b}{||a||.||b||} \tag{5}$$

3.3.6 Research Trend Mapping. The final stage of the proposed framework comprises the mapping of the evolution of scientific trends. To accomplish this, all the inter-period cluster centroids are compared and the most similar neighboring periods (up to two periods further) are connected based on a constant threshold value, which is an empirical threshold value, just over the intra-period similarities. To illustrate the connections among each cluster throughout the periods of time, the Sankey diagram is used in this study.

4 RESULTS

4.1 Document Clustering and Mapping

The method was implemented on 50-dimensional vectors and we noticed that the results from the SIF influenced method lived up to the expectation by providing us with more separable clusters compared to unweighted averaging. This was most noticeable during the cluster number estimation, as the clusters of weighted averaged documents were further apart.

The Wikipedia based labels, which are basically the estimations of Wikipedia AI approaches and applications, based on the similarity of cluster centroids to document vectors, are illustrated at the sample result Tables 2 to 7. [1].

Referring to the aforementioned tables, the overall theme of the topics in the three AI journals can be perceived. The results from Wikipedia Approach Estimation, Wikipedia Application Estimation, and top terms would align well in many cases and it creates sensible topic clusters. For instance, the domination of machine learning (ML) is very natural among AI subjects, which is also the case in the results. In another instance, it is obvious from Table 5 and Table 6 that "Decision support system" peaks during this period, where it was non-existing in the prior periods and also missing in the next period. To validate this claim, the trends of the science can qualitatively be compared during the corresponding periods by searching for decision support system*" and "artificial intelligence" in Scopus (See Fig.5). The appearance of "Automatic target recognition" in conjunction with "Computer vision" from the period of 1995-1999 also aligns with the real world trends of science and the breakthroughs. This pattern of the fields, being sorted next to each other is also interesting and provides further assistance to the interpretation task. Overall, utilizing the proposed method demonstrates the simplicity of the interpretation of topics. It is perceived from the results that Wikipedia applications are the most helpful tags, in contrast to Wikipedia approaches, which only provide minor assistance in understanding the cluster concept.

This also applies to the top words and the yielding analysis, as they are also perceived very helpful in the identification of clusters. As explained at 3.3.5, top terms are very helpful in understanding the cluster. They provide support in deciding the correct cluster tags. Hence, they require expert opinion to form the final cluster label, similar to some prior studies [11]. The alignment between the Wikipedia applications and the top words is visible in many periods. Yet, it should be acknowledged that some experts in the field may disagree on the meaning of the top words and may interpret them differently in comparison to the Wikipedia topics. Therefore, this has been left as it is and expert opinion is not used for this part of the analysis. To facilitate the interpretation of the top words, they've been recorded with the corresponding *tf-idf* score of the term in the cluster of interest. This facilitates the identification of clusters and provides more weights to understand the importance of a term for labeling each cluster. Only sample results are provided in Table 1 due to page constraints.

The evolution mapping of the fields with two different labeling approaches is presented in Fig.6 and Fig.7. This mapping connects the topics extracted throughout all periods. This diagram connects two or three consecutive clusters based on the similarity.

5 CONCLUSION

In this study, we proposed and implemented a framework to extract scientific trends and visualize their evolution in discrete time periods. The study shows that this framework and labeling method facilitates the identification of trends and assist us in understanding the way fields of research are evolving. This became possible through the top term and Wikipedia application labeling methods. We also show that Wikipedia documents can be used to have an estimated embedding location of a field of research or an application in vector space. Yet, Wikipedia approaches are not as useful as Wikipedia application for this case study and purpose. In future works, more advanced clustering methods are planned to be used as an extension to this work, benefiting from deep neural networks in clustering and dynamic embedding and clustering techniques.

[1] Abbreviations: PR: pattern recognition, NLP: natural language processing, ML: machine learning, DSS: decision support system, ES: expert system, KM: knowledge management, CV: computer vision, auto.: automated

REFERENCES

[1] A. Zeng, Z. Shen, J. Zhou, J. Wu, Y. Fan, Y. Wang, and H. E. Stanley, "The science of science: From the perspective of complex systems," *Physics Reports*, vol. 714-715, pp. 1–73, 2017.

[2] J. G. Foster, A. Rzhetsky, and J. A. Evans, "Tradition and innovation in scientist-sâĂŹ research strategies," *American Sociological Review*, vol. 80, no. 5, pp. 875–908, 2015.

[3] S. Fortunato, C. T. Bergstrom, K. Börner, J. A. Evans, D. Helbing, S. Milojević, A. M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi, A. Vespignani, L. Waltman, D. Wang, and A. L. Barabási, "Science of science," *Science*, vol. 359, no. 6379, 2018.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[5] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey," *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15169–15211, 2019.

[6] P. Van den Besselaar and G. Heimeriks, "Mapping research topics using word-reference co-occurrences: A method and an exploratory case study," *Scientometrics*, vol. 68, no. 3, pp. 377–393, 2006.

[7] M. Sedighi, "Application of word co-occurrence analysis method in mapping of the scientific fields (case study: the field of informetrics)," *Library Review*, vol. 65, no. 1/2, pp. 52–64, 2016.

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

[9] W. Zhao, J. Mao, and K. Lu, "Ranking themes on co-word networks: Exploring the relationships among different metrics," *Information Processing & Management*, vol. 54, no. 2, pp. 203–218, 2018.

[10] A. Yang, Q. Lv, F. Chen, D. Wang, Y. Liu, and W. Shi, "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*, vol. 25, p. 643, 2019.

[11] Y. Zhang, H. Chen, J. Lu, and G. Zhang, "Detecting and predicting the topic change of Knowledge-based Systems: A topic-based bibliometric analysis from 1991 to 2016," *Knowledge-Based Systems*, vol. 133, pp. 255–268, 2017.

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

[13] Y. Zhang, G. Zhang, D. Zhu, and J. Lu, "Scientific evolutionary pathways: Identifying and visualizing relationships for scientific topics," *Journal of the Association for Information Science and Technology*, vol. 68, pp. 1925–1939, aug 2017.

[14] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc., 1986.

[15] Y. Zhang, J. Lu, F. Liu, Q. Liu, A. Porter, H. Chen, and G. Zhang, "Does deep learning help topic extraction? a kernel k-means clustering method with word embedding," *Journal of Informetrics*, vol. 12, no. 4, pp. 1099–1117, 2018.

[16] X. Li, Q. Xie, J. Jiang, Y. Zhou, and L. Huang, "Identifying and monitoring the development trends of emerging technologies using patent analysis and twitter data mining: The case of perovskite solar cell technology," *Technological Forecasting and Social Change*, vol. 146, pp. 687–705, 2019.

[17] S. A. Curiskis, B. Drake, T. R. Osborn, and P. J. Kennedy, "An evaluation of document clustering and topic modelling in two online social networks: Twitter and reddit," *Information Processing & Management*, vol. 57, no. 2, p. 102034, 2020.

[18] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*, pp. 1188–1196, 2014.

[19] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, vol. 28, pp. 11–21, 1972.

[20] J. Kim, J. Yoon, E. Park, and S. Choi, "Patent document clustering with deep embeddings," *Scientometrics*, pp. 1–15, 2020.

[21] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint arXiv:1607.01759*, 2016.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[23] M. Steinbach, L. Ertöz, and V. Kumar, "The challenges of clustering high dimensional data," in *New directions in statistical physics*, pp. 273–309, Springer, 2004.

[24] H.-P. Kriegel, P. Kröger, and A. Zimek, "Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 3, no. 1, p. 1, 2009.

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

[26] J. Park, C. Park, J. Kim, M. Cho, and S. Park, "Adc: Advanced document clustering using contextualized representations," *Expert Systems with Applications*, vol. 137, pp. 157–166, 2019.

[27] A. M. Dai, C. Olah, and Q. V. Le, "Document embedding with paragraph vectors," *arXiv preprint arXiv:1507.07998*, 2015.

[28] S. Arora, Y. Liang, and T. Ma, "A simple but tough-to-beat baseline for sentence embeddings," 2017.

[29] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.

[30] V. Saquicela, F. Baculima, G. Orellana, N. Piedra, M. Orellana, and M. Espinoza, "Similarity detection among academic contents through semantic technologies and text mining.," in *IWSW*, pp. 1–12, 2018.

Table 1: Top 7 term tags for the period of 2017-2019.

Tag (TF-IDF score)
* cluster (0.226), clustering (0.194), ba (0.156), twsvm (0.148), support vector machine (0.147), neural network (0.119), si (0.117)
* queen (0.537), kemeny (0.224), top (0.173), bound (0.158), borda (0.153), mining (0.15), item (0.148)
* logic (0.369), semantics (0.218), answer set (0.203), formula (0.179), cp net (0.177), revision (0.152), asp (0.151)
* market (0.257), sale (0.226), firm (0.226), car (0.164), customer (0.157), kidney (0.157), bike (0.157)
* knee (0.319), face recognition (0.253), acl (0.209), gait (0.198), gait pattern (0.176), facial (0.176), survey (0.172)
* planning (0.272), heuristic (0.237), plan (0.201), abstraction (0.181), search (0.177), planner (0.16), monte carlo tree search (0.13)
* sentiment analysis (0.268), survey (0.245), text (0.179), metadata (0.154), area (0.14), indian language (0.133), citation (0.124)
* word (0.271), entity (0.211), sentiment (0.176), vietnamese (0.135), sentiment analysis (0.13), semantic (0.124), target (0.122)
* voting (0.233), voter (0.218), cost (0.16), mirl (0.15), player (0.142), good (0.141), preference (0.139)
* inconsistency (0.231), semantics (0.156), attack (0.153), belief (0.153), argument (0.143), graph (0.139), argumentation framework (0.136)
robot (0.401), team (0.217), trust (0.17), teammate (0.139), belief (0.121), revision (0.12), norm (0.112)

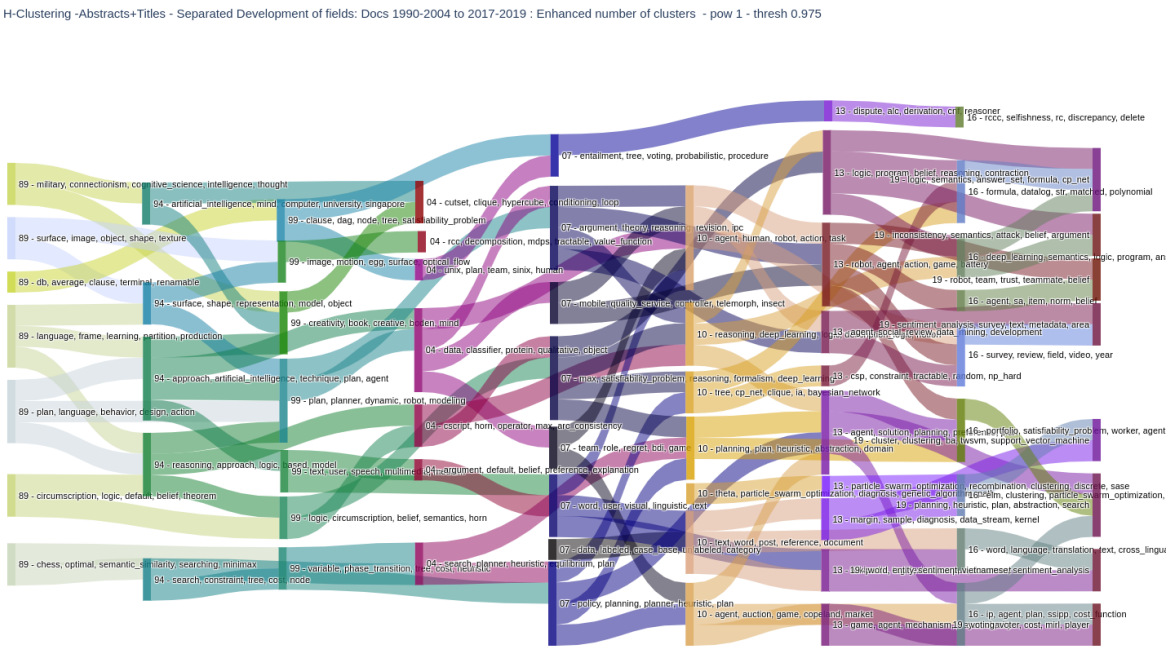

Figure 6: Sankey diagram of the research clusters with top words as labels. The initial digits are referring to the ending year of period.

Table 2: Wikipedia based approaches and applications (topics) for the period of 2000-2004.

Wiki Approach Est.	Wiki Application Est.
Probability & Chaos theory	Nonlinear control & auto. planning and scheduling
ML & Behavior based AI	Intelligent agent & ML
ML & Fuzzy systems	ML & auto. planning and scheduling
Probability & ML	auto. planning and scheduling & auto. reasoning
ML & Behavior based AI	Computer audition & NLP
Fuzzy systems & ML	CV and subfields & Automatic target recognition
Early cybernetics and brain simulation & Behavior based AI	Computational creativity & auto. reasoning

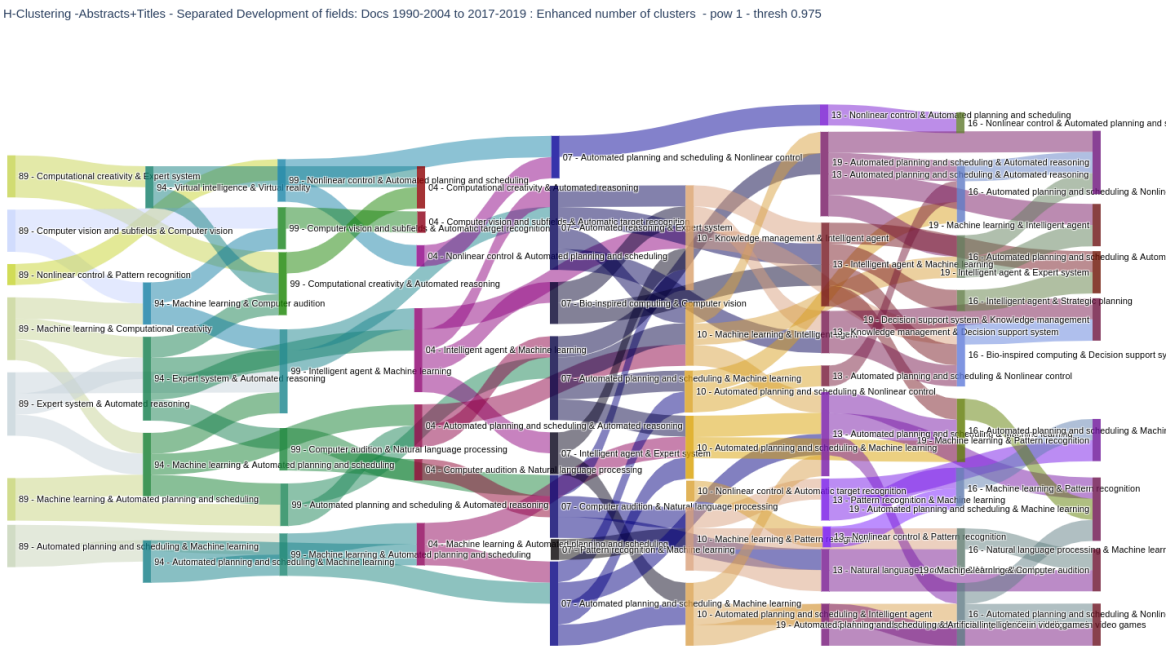

Figure 7: Sankey diagram with Wikipedia applications as labels. The initial digits are referring to the ending year of period.

Table 3: Wikipedia based approaches and applications (topics) for the period of 2005-2007.

Wiki Approach Est.	Wiki Application Est.
Probability & Chaos theory	auto. planning and scheduling & Nonlinear control
ML & Fuzzy systems	auto. planning and scheduling & ML
ML & Behavior based AI	Computer audition & NLP
ML & Behavior based AI	auto. reasoning & ES
ML & Probability	auto. planning and scheduling & ML
Behavior based AI & ML	Bio-inspired computing & CV
ML & Behavior based AI	Intelligent agent & ES
ML & Evolutionary computation	PR & ML

Table 4: Wikipedia based approaches and applications (topics) for the period of 2008-2010.

Wiki Approach Est.	Wiki Application Est.
ML & Fuzzy systems	ML & PR
Behavior based AI & ML	KM & Intelligent agent
Probability & Chaos theory	Nonlinear control & auto. planning and scheduling
Fuzzy systems & Chaos theory	auto. planning and scheduling & Intelligent agent
ML & Probability	ML & Intelligent agent
Fuzzy systems & ML	Nonlinear control & Automatic target recognition
Chaos theory & Probability	auto. planning and scheduling & Nonlinear control
Evolutionary computation & Early cybernetics and brain simulation	Hybrid intelligent system & Bio-inspired computing
ML & Fuzzy systems	auto. planning and scheduling & ML

Table 5: Wikipedia based approaches and applications (topics) for the period of 2011-2013.

Wiki Approach Est.	Wiki Application Est.
ML & Behavior based AI	Intelligent agent & ML
Chaos theory & Probability	auto. planning and scheduling & Nonlinear control
Early cybernetics and brain simulation & Behavior based AI	KM & Decision support system
Probability & ML	auto. planning and scheduling & auto. reasoning
Chaos theory & Probability	auto. planning and scheduling & AI in video games
ML & Behavior based AI	NLP & ML
ML & Fuzzy systems	auto. planning and scheduling & ML
ML & Fuzzy systems	PR & Intelligent control
Probability & Chaos theory	Nonlinear control & auto. planning and scheduling
ML & Chaos theory	PR & ML
Evolutionary computation & ML	Nonlinear control & PR

Table 6: Wikipedia based approaches and applications (topics) for the period of 2014-2016.

Wiki Approach Est.	Wiki Application Est.
Behavior based AI & Early cybernetics and brain simulation	Bio-inspired computing & Decision support system
Probability & Chaos theory	auto. planning and scheduling & Nonlinear control
Evolutionary computation & AI	AI & PR
AI & Chaos theory	CV and subfields & Automatic target recognition
Fuzzy systems & Chaos theory	auto. planning and scheduling & Nonlinear control
AI & Behavior based AI	NLP & AI
AI & Probability	auto. planning and scheduling & auto. reasoning
AI & Probability	Intelligent agent & Strategic planning
Probability & Chaos theory	Nonlinear control & auto. planning and scheduling
AI & Fuzzy systems	PR & Nonlinear control
AI & Fuzzy systems	auto. planning and scheduling & AI

Table 7: Wikipedia based approaches and applications (topics) for the period of 2017-2019.

Wiki Approach Est.	Wiki Application Est.
ML & Fuzzy systems	ML & PR
Probability & Chaos theory	PR & Nonlinear control
Probability & Chaos theory	auto. planning and scheduling & auto. reasoning
Fuzzy systems & Behavior based AI	Automation & Vehicle infrastructure integration
Behavior based AI & ML	CV & Computer audition
ML & Fuzzy systems	auto. planning and scheduling & ML
Early cybernetics and brain simulation & ML	DSS& KM
ML & Behavior based AI	ML & Computer audition
Probability & Chaos theory	auto. planning and scheduling & AI in video games
ML & Probability	ML & Intelligent agent
ML & Behavior based AI	Intelligent agent & ES