ACM/IEEE Joint Conference on Digital Libraries (JCDL 2020)
Wuhan, Hubei, P. R. China
August 1 - 5, 2020

Document Clustering and Labeling for Research Trend Extraction and Evolution Mapping

1st Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2020)

Presenter: Sahand Vahidnia
Coautors: Alireza Abbasi & Hussein A. Abbass
{s.vahidnia , a.abbasi , h.abbass}@unsw.edu.au
UNSW CANBERRA

01/08/2020

Introduction

Understanding and predicting future discoveries and scientific achievements is an emerging field of research, which involves scientists, businesses, and even governments.

This topic falls under the emerging field of Science of Science (SciSci) which aims to understand, quantify and predict scientific research dynamics and the drivers of the dynamics in different forms such as the birth and death of scientific fields and their subfields; that can be identified by tracking the changes of research trends and dynamics.

Objective & Outline

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.

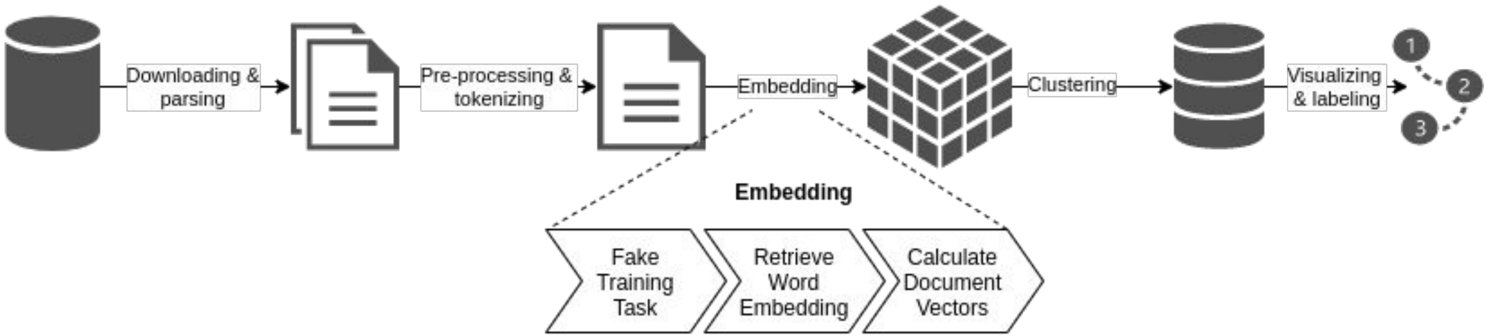
Our proposed framework is then applied in a case study to understand the path of scientific evolution in artificial intelligence. In this study, we show how the trends and topics in science can be extracted using document vectors and extraction of context.

The Literature

- Co-word analysis & topic modeling
  - [6],[7],[8],[9],[10],[11]
- Word and document embedding and clustering
  - [12],[13],[15],[16],[17],[18]

- Embedding methods are superior to traditional methods like TF-IDF for clustering tasks.
- A framework to detect, track and visualize the trends in alluvial like diagrams is out of focus

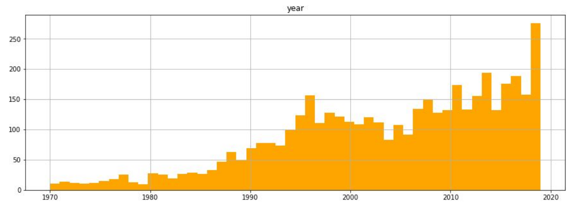
Methodology



Methodology

1) Data Collection

- Dataset A (Word embedding model training data):
  - Abstracts and titles
  - Scopus search in titles, abstracts, and keywords with ``artificial intelligence'' query
  - Yielding 310k records
- Dataset B (Case study and analysis data):
  - Abstracts and titles
  - 3 mainstream journals:
    - ``Artificial Intelligence'' (2575 records)
    - ``Artificial Intelligence Review'' (890 records)
    - ``Journal of Artificial Intelligence Research'' (1006 records)



Methodology

2) Preprocessing

- Common data pre-processing: Carried out on both datasets, including data cleaning, removal of common abbreviations and noun level lemmatization.
- Analysis data preprocessing: (Carried out on dataset B.)
  - Removal of stop words.
  - N-gram keyword tagging and creation of auxiliary labeling dataset.
  - Splitting data to temporal periods: [1970,1989] , [1990,1994], [1995,1999], [2000,2004], [2005,2007], [2008,2010], [2011,2013], [2014,2016], and [2017,2019].

Methodology

3) Word Embedding

- Represent the data in vector space.
- FastText is used to extract word vectors.
- Dataset A is used to train the model.
- Embeddings are in 50 dimensions.
- No further dimensionality reduction is used.

Methodology

4) Document Embedding

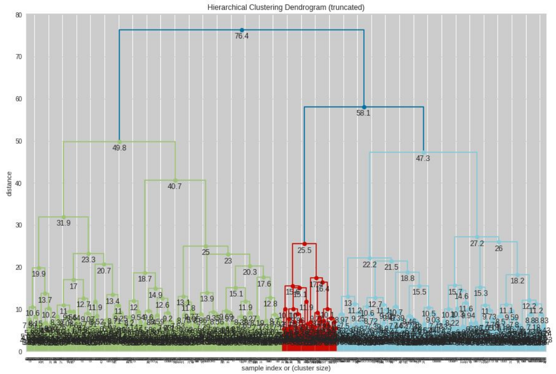
- Simple document averaging
- SIF [28]

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

$$wv(t) = w(t) * v(t)$$

5) Document Clustering

- Hierarchical agglomerative clustering.
- Assist in number of clusters by dendrogram.



Methodology

6) Cluster Labeling

- Document keywords

$$score(t,c) = tf(t,c) * icf(t)$$

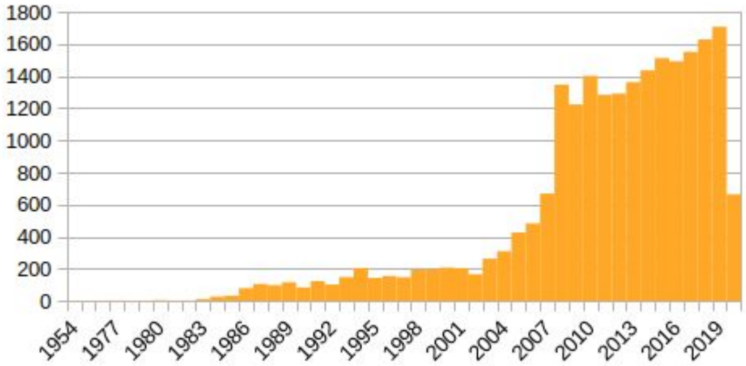
$$icf(t) = log\frac{(1+n)}{(1+cf(t))} + 1$$

- Wikipedia labels
    - Applications
    - Approaches

7) Research Trend Mapping

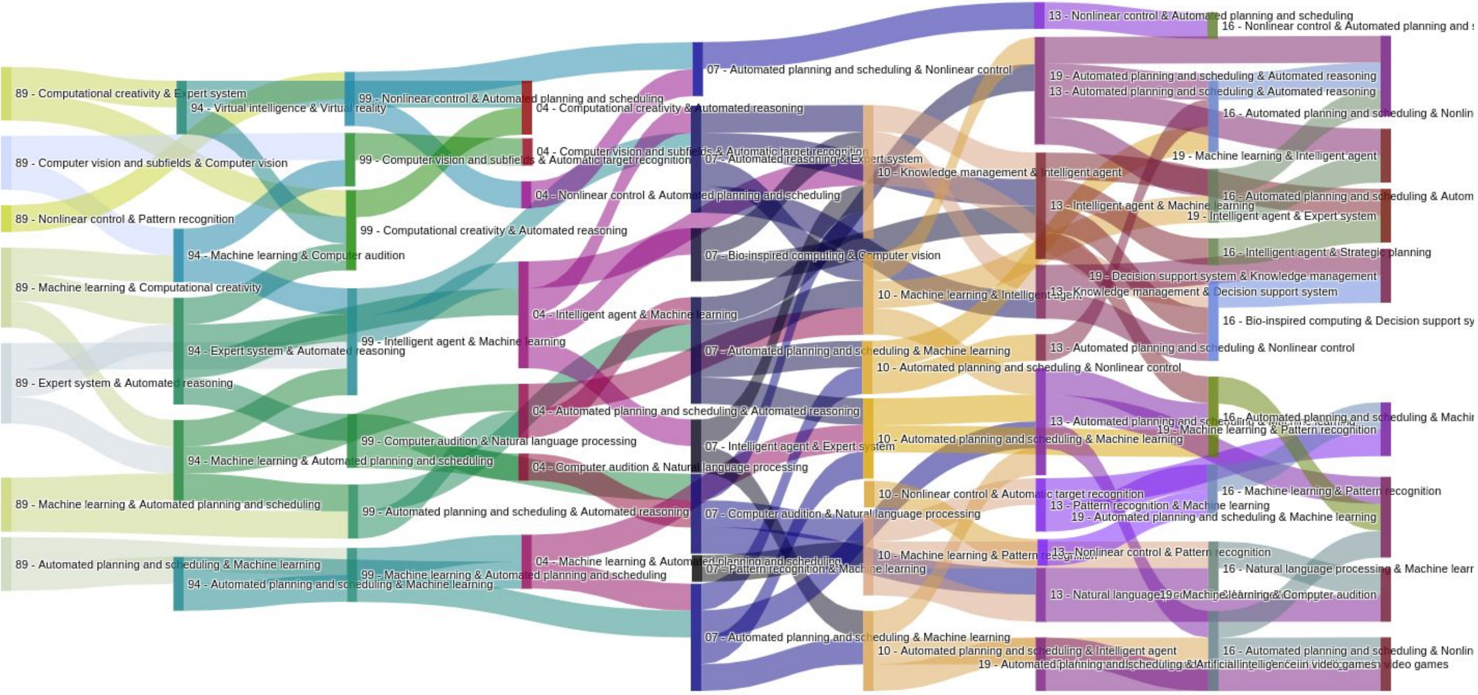
- The final stage of the proposed framework comprises the mapping of the evolution of scientific trends.

Results (1/4)



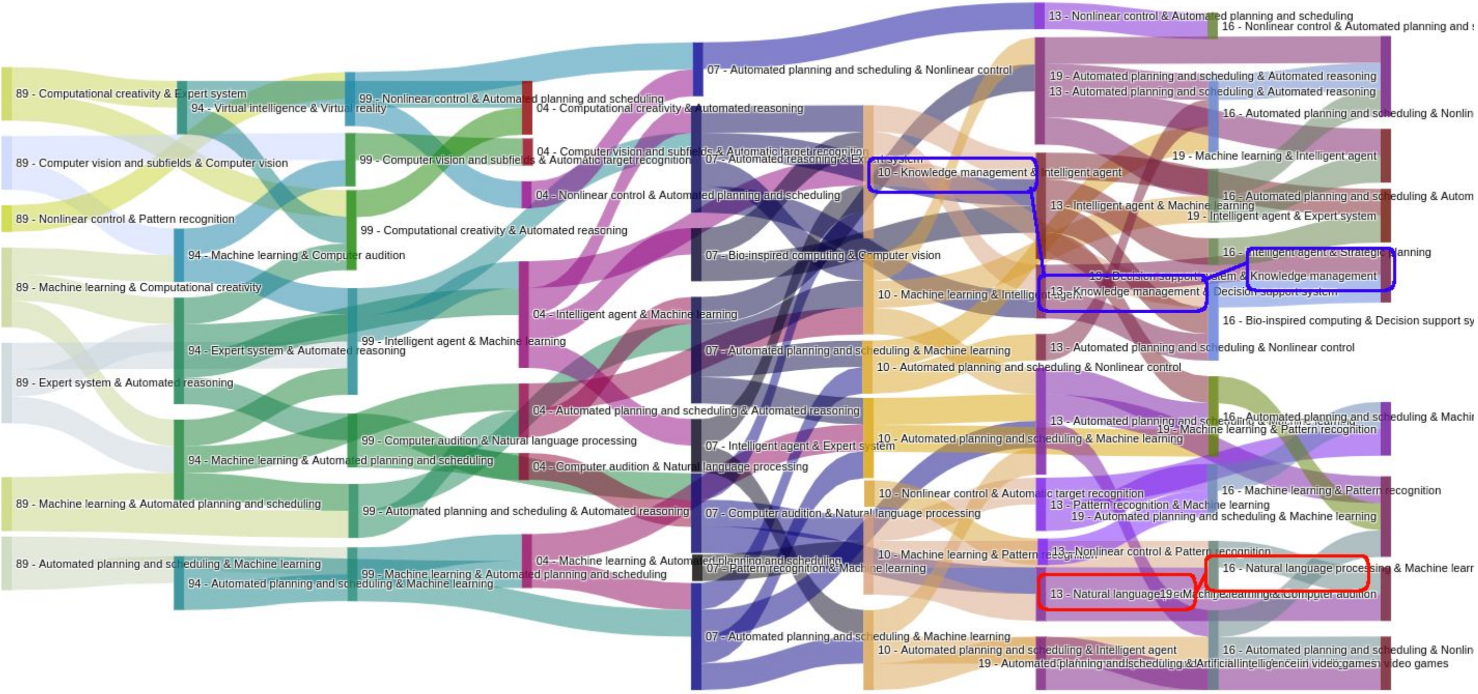
Wiki Application Est. 2011-2013	Wiki Application Est. 2014-2016
Intelligent agent & ML	Bio-inspired computing & Decision support system
auto. planning and scheduling & Nonlinear control	auto. planning and scheduling & Nonlinear control
KM & Decision support system	AI & PR
auto. planning and scheduling & auto. reasoning	CV and subfields & Automatic target recognition
auto. planning and scheduling & AI in video games	auto. planning and scheduling & Nonlinear control
NLP & ML	NLP & AI
auto. planning and scheduling & ML	auto. planning and scheduling & auto. reasoning
PR & Intelligent control	Intelligent agent & Strategic planning
Nonlinear control & auto. planning and scheduling	Nonlinear control & auto. planning and scheduling
PR & ML	PR & Nonlinear control
Nonlinear control & PR	auto. planning and scheduling & AI

Results (2/4)

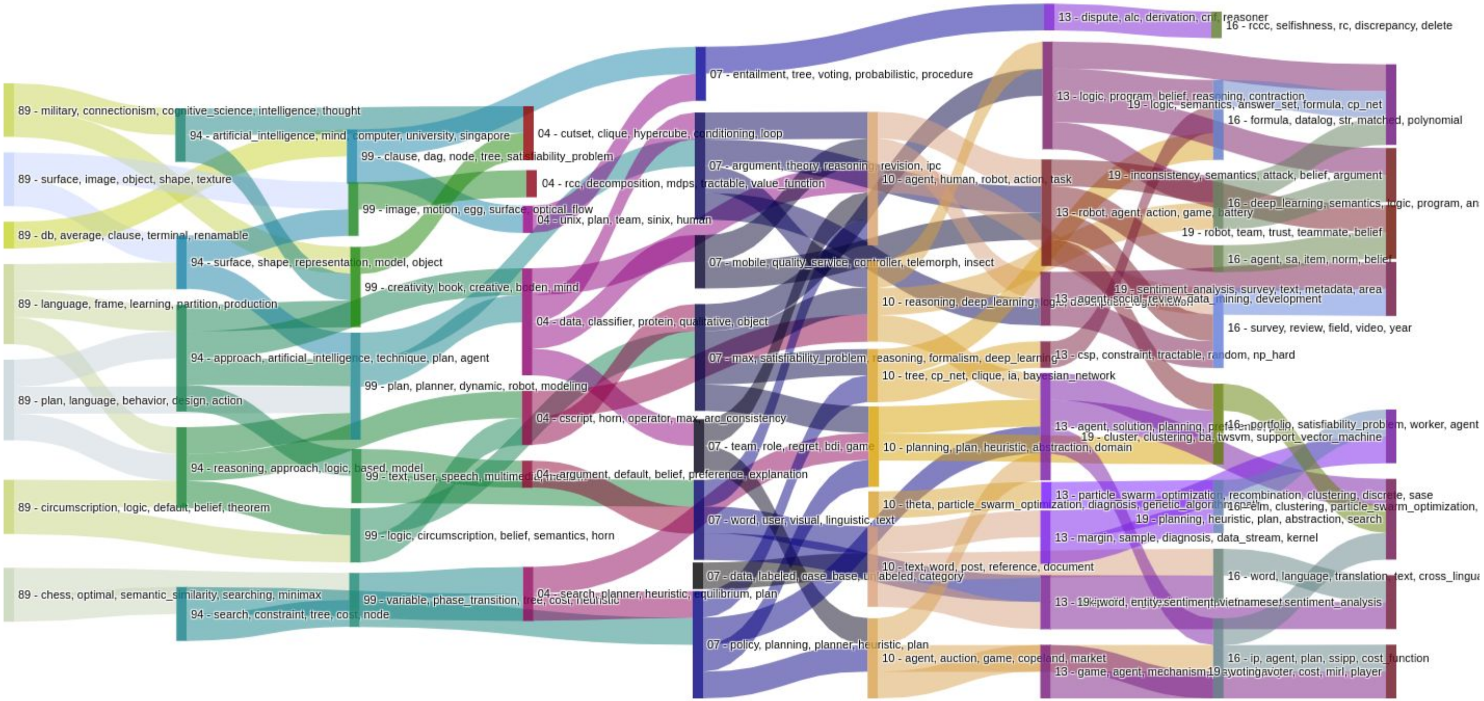


89 - Computational creativity & Expert system
89 - Computer vision and subfields & Computer vision
89 - Nonlinear control & Pattern recognition
89 - Machine learning & Computational creativity
89 - Expert system & Automated reasoning
89 - Machine learning & Automated planning and scheduling
89 - Automated planning and scheduling & Machine learning
94 - Virtual intelligence & Virtual reality
94 - Machine learning & Computer audition
94 - Expert system & Automated reasoning
94 - Machine learning & Automated planning and scheduling
94 - Automated planning and scheduling & Machine learning
99 - Nonlinear control & Automated planning and scheduling
99 - Computer vision and subfields & Automatic target recognition
99 - Computational creativity & Automated reasoning
99 - Intelligent agent & Machine learning
99 - Computer audition & Natural language processing
99 - Automated planning and scheduling & Automated reasoning
99 - Machine learning & Automated planning and scheduling
04 - Computational creativity & Automated reasoning
04 - Computer vision and subfields & Automatic target recognition
04 - Nonlinear control & Automated planning and scheduling
04 - Intelligent agent & Machine learning
04 - Automated planning and scheduling & Automated reasoning
04 - Computer audition & Natural language processing
04 - Machine learning & Automated planning and scheduling
07 - Automated planning and scheduling & Nonlinear control
07 - Automated reasoning & Expert system
07 - Bio-inspired computing & Computer vision
07 - Automated planning and scheduling & Machine learning
07 - Intelligent agent & Expert system
07 - Computer audition & Natural language processing
07 - Pattern recognition & Machine learning
07 - Automated planning and scheduling & Machine learning
10 - Knowledge management & Intelligent agent
10 - Machine learning & Intelligent agent
10 - Automated planning and scheduling & Nonlinear control
10 - Automated planning and scheduling & Machine learning
10 - Nonlinear control & Automatic target recognition
10 - Machine learning & Pattern recognition
10 - Automated planning and scheduling & Intelligent agent
13 - Nonlinear control & Automated planning and scheduling
13 - Automated planning and scheduling & Automated reasoning
13 - Intelligent agent & Machine learning
13 - Knowledge management & Decision support system
13 - Automated planning and scheduling & Nonlinear control
13 - Automated planning and scheduling & Machine learning
13 - Pattern recognition & Machine learning
13 - Nonlinear control & Pattern recognition
19 - Automated planning and scheduling & Automated reasoning
19 - Machine learning & Intelligent agent
19 - Intelligent agent & Expert system
19 - Decision support system & Knowledge management
19 - Machine learning & Pattern recognition
19 - Automated planning and scheduling & Machine learning
19 - Machine learning & Computer audition
16 - Nonlinear control & Automated planning and
16 - Automated planning and scheduling & Nonlin
16 - Automated planning and scheduling & Autom
16 - Intelligent agent & Strategic planning
16 - Bio-inspired computing & Decision support sy
16 - Automated planning and scheduling & Machir
16 - Machine learning & Pattern recognition
16 - Natural language processing & Machine learr
16 - Automated planning and scheduling & Nonlin

Results (2/4)


89 - Computational creativity & Expert system
89 - Computer vision and subfields & Computer vision
89 - Nonlinear control & Pattern recognition
89 - Machine learning & Computational creativity
89 - Expert system & Automated reasoning
89 - Machine learning & Automated planning and scheduling
89 - Automated planning and scheduling & Machine learning
94 - Virtual intelligence & Virtual reality
94 - Machine learning & Computer audition
94 - Expert system & Automated reasoning
94 - Machine learning & Automated planning and scheduling
94 - Automated planning and scheduling & Machine learning
99 - Nonlinear control & Automated planning and scheduling
99 - Computer vision and subfields & Automatic target recognition
99 - Computational creativity & Automated reasoning
99 - Intelligent agent & Machine learning
99 - Computer audition & Natural language processing
99 - Automated planning and scheduling & Automated reasoning
99 - Machine learning & Automated planning and scheduling
04 - Computational creativity & Automated reasoning
04 - Computer vision and subfields & Automatic target recognition
04 - Nonlinear control & Automated planning and scheduling
04 - Intelligent agent & Machine learning
04 - Automated planning and scheduling & Automated reasoning
04 - Computer audition & Natural language processing
04 - Machine learning & Automated planning and scheduling
07 - Automated planning and scheduling & Nonlinear control
07 - Automated reasoning & Expert system
07 - Bio-inspired computing & Computer vision
07 - Automated planning and scheduling & Machine learning
07 - Intelligent agent & Expert system
07 - Computer audition & Natural language processing
07 - Pattern recognition & Machine learning
07 - Automated planning and scheduling & Machine learning
10 - Knowledge management & Intelligent agent
10 - Machine learning & Intelligent agent
10 - Automated planning and scheduling & Nonlinear control
10 - Automated planning and scheduling & Machine learning
10 - Nonlinear control & Automatic target recognition
10 - Machine learning & Pattern recognition
10 - Automated planning and scheduling & Intelligent agent
13 - Nonlinear control & Automated planning and scheduling
13 - Automated planning and scheduling & Automated reasoning
13 - Intelligent agent & Machine learning
13 - Knowledge management & Decision support system
13 - Automated planning and scheduling & Nonlinear control
13 - Pattern recognition & Machine learning
13 - Nonlinear control & Pattern recognition
19 - Automated planning and scheduling & Automated reasoning
19 - Machine learning & Intelligent agent
19 - Intelligent agent & Expert system
19 - Machine learning & Pattern recognition
19 - Automated planning and scheduling & Machine learning
16 - Automated planning and scheduling & Nonlin
16 - Bio-inspired computing & Decision support sy
16 - Machine learning & Pattern recognition
16 - Natural language processing & Machine learr

Results (3/4)



89 - military, connectionism, cognitive_science, intelligence, thought
89 - surface, image, object, shape, texture
89 - db, average, clause, terminal, renamable
89 - language, frame, learning, partition, production
89 - plan, language, behavior, design, action
89 - circumscription, logic, default, belief, theorem
89 - chess, optimal, semantic_similarity, searching, minimax
94 - artificial_intelligence, mind, computer, university, singapore
94 - surface, shape, representation, model, object
94 - approach, artificial_intelligence, technique, plan, agent
94 - reasoning, approach, logic, based, model
94 - search, constraint, tree, cost, node
99 - clause, dag, node, tree, satisfiability_problem
99 - image, motion, egg, surface, optical_flow
99 - creativity, book, creative, boden, mind
99 - plan, planner, dynamic, robot, modeling
99 - logic, circumscription, belief, semantics, horn
99 - variable, phase_transition, tree, cost, heuristic
04 - cutset, clique, hypercube, conditioning, loop
04 - rcc, decomposition, mdps, tractable, value_function
04 - unix, plan, team, sinix, human
04 - data, classifier, protein, qualitative, object
04 - cscript, horn, operator, max, arc_consistency
04 - argument, default, belief, preference, explanation
04 - search, planner, heuristic, equilibrium, plan
07 - entailment, tree, voting, probabilistic, procedure
07 - argument, theory, reasoning, revision, ipc
07 - mobile, quality_service, controller, telemorph, insect
07 - max, satisfiability_problem, reasoning, formalism, deep_learning
07 - team, role, regret, bdi, game
07 - word, user, visual, linguistic, text
07 - data, labeled, case_base, unlabeled, category
07 - policy, planning, planner, heuristic, plan
10 - agent, human, robot, action, task
10 - reasoning, deep_learning, logic, description_logic, notion
10 - tree, cp_net, clique, ia, bayesian_network
10 - planning, plan, heuristic, abstraction, domain
10 - theta, particle_swarm_optimization, diagnosis, genetic_algorithm, elm
10 - text, word, post, reference, document
10 - agent, auction, game, copeland, market
13 - dispute, alc, derivation, cnf, reasoner
13 - logic, program, belief, reasoning, contraction
13 - robot, agent, action, game, battery
13 - agent, social, review, data_mining, development
13 - csp, constraint, tractable, random, np_hard
13 - agent, solution, planning, preference, plan
13 - particle_swarm_optimization, recombination, clustering, discrete, sase
13 - margin, sample, diagnosis, data_stream, kernel
13 - game, agent, mechanism
16 - rccc, selfishness, rc, discrepancy, delete
16 - formula, datalog, str, matched, polynomial
16 - deep_learning, semantics, logic, program, an
16 - agent, sa, item, norm, belief
16 - survey, review, field, video, year
16 - portfolio, satisfiability_problem, worker, agent
16 - elm, clustering, particle_swarm_optimization,
16 - word, language, translation, text, cross_lingua
16 - ip, agent, plan, ssipp, cost_function
19 - logic, semantics, answer_set, formula, cp_net
19 - inconsistency, semantics, attack, belief, argument
19 - robot, team, trust, teammate, belief
19 - sentiment_analysis, survey, text, metadata, area
19 - cluster, clustering, ba, twsvm, support_vector_machine
19 - planning, heuristic, plan, abstraction, search
vietnamese, sentiment_analysis
voter, cost, mirl, player

Results (4/4)

	Tag (TF-IDF score) - 2017-2019
1	* cluster (0.226), clustering (0.194), ba (0.156), twsvm (0.148), support vector machine (0.147), neural network (0.119), si (0.117)
2	* queen (0.537), kemeny (0.224), top (0.173), bound (0.158), borda (0.153), mining (0.15), item (0.148)
3	* logic (0.369), semantics (0.218), answer set (0.203), formula (0.179), cp net (0.177), revision (0.152), asp (0.151)
4	* market (0.257), sale (0.226), firm (0.226), car (0.164), customer (0.157), kidney (0.157), bike (0.157)
5	* knee (0.319), face recognition (0.253), acl (0.209), gait (0.198), gait pattern (0.176), facial (0.176), survey (0.172)
6	* planning (0.272), heuristic (0.237), plan (0.201), abstraction (0.181), search (0.177), planner (0.16), monte carlo tree search (0.13)
7	* sentiment analysis (0.268), survey (0.245), text (0.179), metadata (0.154), area (0.14), indian language (0.133), citation (0.124)
8	* word (0.271), entity (0.211), sentiment (0.176), vietnamese (0.135), sentiment analysis (0.13), semantic (0.124), target (0.122)
9	* voting (0.233), voter (0.218), cost (0.16), mirl (0.15), player (0.142), good (0.141), preference (0.139)
10	* inconsistency (0.231), semantics (0.156), attack (0.153), belief (0.153), argument (0.143), graph (0.139), argumentation framework (0.136)
11	* robot (0.401), team (0.217), trust (0.17), teammate (0.139), belief (0.121), revision (0.12), norm (0.112)

Conclusion

- 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.
- Wikipedia documents can be used to have an estimated embedding location of a field of research or an application in vector space.
- 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. Additionally, labeling can benefit from the vector space similarities to enhance TF-IDF weights.

Q&A

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 scientistsâĂŹ 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 wordreference 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,” Medicalscience monitor: international medical journal of experimental and clinical research, vol. 25, p. 643, 2019.

References

- [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.

References

- [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.