What's the trend of hot areas in Computer Science field?

Fan Xie *, Rui Jin *, Yijun Tian *
New York University
Courant Institute of Mathematical Sciences
fx349, rj1417, yt1506@nyu.edu

***Abstract*—Forecasting research trends has always been a dream of researchers, scientists, investors and those who want to step into a new area. However, one can not predict the next hot area before he know the trend of past hot areas. This incurs the most interesting question - what's the trend of hot areas in Computer Science field? Most of the trend prediction are mainly based on some experts and there always existing some difference between each one, which also ignore the valuable data part. Thus, how to find the trend of hot areas in the perspective of data is quite a crucial yet challenging problem. In this paper, we propose a novel technique that analyze the trend of hot areas reasonably. The experiment results show that our work performs well at showing the trend of hot areas in computer science area.**

***Index Terms*—Trend analysis, arXiv, computer science**

I. Introduction

Topics in hot research area can attract people's attention. Question like "What are the next hot area in computer science research?" appeared on Quora every year. People are enthusiastic about asking and answering these questions. However, with so many areas like Artificial Intelligence, Hardware Architecture, Databases, Graphics in Computer Science field, how to find the next hot area reasonable and precisely is quite a crucial yet challenging problem. Currently, most of the trend prediction are mainly based on some experts and there always existing some difference between each one. The lack of consistent metric also makes the prediction results hard to evaluate. Therefore, predicting the next hot area by analyzing the trend of past hot areas in the perspective of data is natural, reliable and credible.

To find out the trend of hot areas is of great value. Students who will choose their major want to know which subject is the most promising. Researchers want to know which subject attract the most attention. In other words, publication with hot topic may not only get a better chance of acceptance into conferences or journals, but also may have influence on the future research interest, investment orientation and employment market. Therefore, understanding future research trends is of great value. However, as the number of papers published annually in each area increases dramatically, it becomes difficult to distinguish between topics that have long-term scientific implications.

In this paper, we analyzed the hot topics in the past years and developed an novel methodology to analyze the trend of hot areas.

* indicates equally contributed.

The main contributions of this paper are summarized as follows:

- we propose a novel schema that automatically analyze the trend of hot areas in computer science field by analyzing and utilizing the past years data.
- We derive a new method to deal with the trend analyzing problem, based on the paper-based features, author-based features, venue-based features, as well as an approach to combine them with a particular weighting factor.
- we combine the result and plot the figure to visualize the result, which makes the exploration of the trend in multiple area easier.

II. Motivation

Computer Science field researchers are able to study the changing trend of the whole industry. Newcomers who want to step into the computer field can get a taste whats the most attractive area currently within the field. Especially for Venture capital, whats the next hot area is the most attractive question. The hot area point out the research interests of researchers, which is also a significant indicators to the degree of attention received from scientific communities. Therefore, Venture capitalist can easily track hot topics and make appropriate assessments and investments.

III. Related Work

A. Trend analysis in general

E.W. Saad, D.V. Prokhorov and D.C. Wunsch [10] focused on limiting false alarms and exploited time delay, recurrent, and probabilistic neural networks to predict stock trends. Gabriel Pui Cheong Fung, Jeffrey Xu Yu and Wai Lam [2] presented a system that predicts the changes of stock trend by analyzing the influence of non-quantifiableinformation (news articles). Jung-Hua Wang and Jia-Yann Leu [5] utilized recurrent neural network trained by using features extracted from ARIMA analyses to forecast mid-term price trend in Taiwan stock market. Jos Carlos Oliveira Santos and Sergio Pazzini da Silva Matos [11] presented an infodemiology study that evaluates the use of Twitter messages and search engine query logs to estimate and predict the incidence rate of influenza like illness in Portugal. Dominique Lord and Bhagwant N. Persaud [6] presented of a generalized estimating equations (GEE) procedure to develop an APM that incorporates trend in accident data.

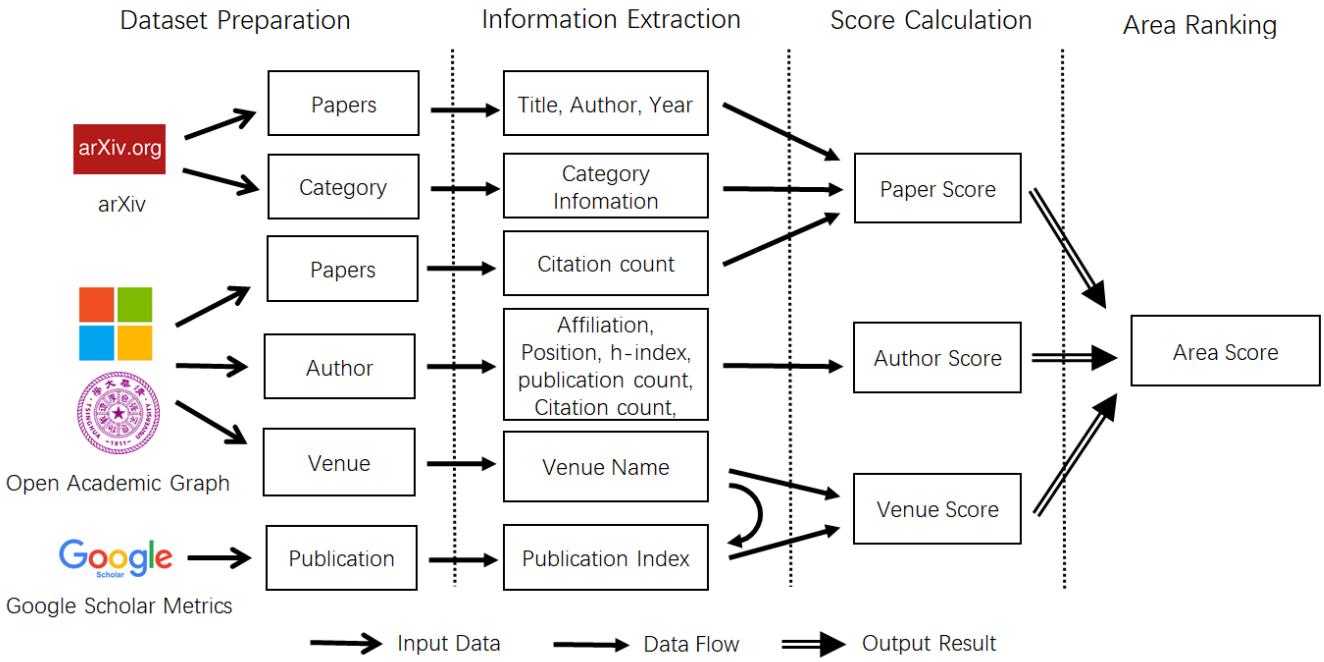

Fig. 1. Pipeline of prediction of next hot area in Computer Science field.

B. Trend analysis in academic

Satyam Mukherjee, Daniel M. Romero, Ben Jones and Brian Uzzi [7] parameterized the age distribution of works references and revealed three links between the age of prior knowledge and hit papers and patents to identify prospectively high-impact science. Vinodkumar Prabhakaran, William L. Hamilton, Daniel A. McFarland and Daniel Jurafsky [8] investigated the role of rhetorical framing to identify scientific topics that will increase or decline over time. Naoki Shibata, Yuya Kajikawa, Yoshiyuki Takeda and Katsumori Matsushima [12] performed topological measures to detecting branching innovation in the citation network of scientific publications. Qi He, et al. [3] proposed an iterative topic evolution learning framework detect trend by adapting the Latent Dirichlet Allocation model to the citation network Arjun Duvvuru, Sagar Kamarthi and Sivarit Sultornsanee [1] analyzed and represented keywords appearing in scholarly articles to uncover trends in an area of research. Jinghao Zhao, Hao Wu, Fengyu Deng, Wentian Bao, Wencheng Tang, Luoyi Fu and Xinbing Wang [15] considered both inter- and intra-topical influence to forecast the next hot topics in Scholarly Fields. Jose L. Hurtado, Ankur Agarwal and Xingquan Zhu [4] used association analysis to automatically discover topics from a set of text documents and forecast their evolving trend in a near future. Tieyun Qian, Qing Li, Bing Liu, Hui Xiong, Jaideep Srivastava and Phillip C.Y. Sheu [9] proposed a model based on the relation of papers in one topic and predicted the core-groups life circle.

IV. Design and Implementation

A. Design Details

1) Areas in arXiv: The Computer Science section in arXiv are the Computing Research Repository (CoRR), which was established in 1998 through a partnership of the Association for Computing Machinery, the Networked Computer Science Technical Reference Library, and arXiv.

The area within arXiv can be found in Table II

2) Score of paper calculation: In arXiv, we first extract paper information and the related category information this paper belongs to. Among the paper information, which contain title, author and year, we combine it with the Open Academic Graph to calculate the paper score under the specific category. Here we incorporate the number of citation this paper has received.

The simplest measure for a papers influence is the number of citations it has received. But plain citation counts may be misleading. They may vary depending on the research field and the date of the publication. Instead, citation counts can be normalized by comparing only papers in the same research fields and adjusting citation count scores according to other papers score.

$$S_{paper} = Normalize(citation\_count) \tag{1}$$

where the S_{paper} represents the score of paper and $citation\_count$ indicated the number of citation this paper has received.

TABLE I
TABLE 1: COMPUTER SCIENCE AREA FROM ARXIV.

Artificial Intelligence	Computation and Language	Computational Complexity
Computational Engineering, Finance, and Science	Computational Geometry	Computer Science and Game Theory
Computer Vision and Pattern Recognition	Computers and Society	Cryptography and Security
Data Structures and Algorithms	Databases	Digital Libraries
Discrete Mathematics	Distributed, Parallel, and Cluster Computing	Emerging Technologies
Formal Languages and Automata Theory	General Literature	Graphics
Hardware Architecture	Human-Computer Interaction	Information Retrieval
Information Theory	Logic in Computer Science	Machine Learning
Mathematical Software	Multiagent Systems	Multimedia
Networking and Internet Architecture	Neural and Evolutionary Computing	Numerical Analysis
Operating Systems	Other Computer Science	Performance
Programming Languages	Robotics	Social and Information Networks
Software Engineering	Sound	Symbolic Computation
Systems and Control		

TABLE II
TABLE 2: TOP 10 HOT AREAS IN COMPUTER SCIENCE FIELD/

Rank	Abbreviation	Hot Area Name
1	CS.LG	Machine Learning
2	CS.DS	Data Structure and Algorithm
3	CS.IT	Information Theory
4	CS.DM	Discrete Mathematics
5	CS.CG	Computational Geometry
6	CS.CC	Computational Complexity
7	CS.GT	Game Theory
8	CS.DC	Distributed Parallel, and Cluster Computing
9	CS.SI	Social and Information Networks
10	CS.DB	Databases

3) Score of author calculation:

$$S_{author} = 0.3 \times S_{affiliation} + 0.7 \times S_{publication} \quad (2)$$

$$\begin{aligned} S_{affiliation} &= 0.5 \times S_{affiliation\_rank} \\ &\quad + 0.5 \times S_{position} \end{aligned} \quad (3)$$

$$\begin{aligned} S_{publication} &= 0.3 \times S_{author\_h\_index} \\ &\quad + 0.3 \times S_{publication\_count} \\ &\quad + 0.3 \times S_{citation\_count} \end{aligned} \quad (4)$$

4) Score of venue calculation: In order to take the impact factor of different venue into account. we include the venue information of each paper within the specific given area, which is defined as

$$S_{venue} = 0.5 \times S_{h5\_index} + 0.5 \times S_{h5\_median\_index} \quad (5)$$

5) Score of area calculation: After calculate the score of paper S_{paper}, score of author S_{author} and score of area S_{area}, we combine them to generate the final score of area S_{area}. Then, by sorting S_{area}, the hot area ranking result is generated. The calculation of S_{area} is defined in

$$S_{area} = \sum_{i=1}^{paperNum_{area}} S_{paper_i} + S_{author_i} + S_{venue_i} \quad (6)$$

where S_{area} indicates the score of area and $paperNum_{area}$ represents the total number of paper under $area$.

B. Description of Datasets

1) arxiv: arxiv is a document submission and retrieval system that is heavily used by computer science communities. It has become the primary means of communicating cutting-edge manuscripts on current and ongoing research, providing Open access to 1,574,046 e-prints in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics. Almost all scientific papers are self-archived on the arXiv repository. Manuscripts are often submitted to the arXiv before they are published by more traditional means. In some cases they may never be submitted or published elsewhere, and in others, arXiv-hosted manuscripts are used as the submission channel to traditional publishers such as the American Physical Society, and newer forms of publication such as the Journal for High Energy Physics and overlay journals. arxiv API allows application developers to access all of the arXiv data, search and linking facilities with an easy-to-use programmatic interface. The API can be called with an HTTP request of type GET or POST.

2) Open Academic Graph: With the aim of creating a shared, open and expanding knowledge graph of research and education-focused entities and relationships, Open Academic Graph (OAG) [13], [14] is provided by Open Academic Society. In the first version proposed in 2017, it contains 166,192,182 papers from Microsoft Academic Graph (MAG) and 154,771,162 papers from AMiner and generated 64,639,608 linking (matching) relations between the two graphs. In the second version of Open Academic Graph (OAG v2), which is proposed in 2019, author, venue and newer publication data and the corresponding matchings are incorporated. Besides, since the two large graphs (MAG and AMiner) are both evolving, MAG snapshot in November 2018 and AMiner snapshot in July 2018 or January 2019 was included in OAG v2. The data in OAG v2 can be split into three table: OAG paper table, OAG author and OAG venue table.

V. RESULTS

Our team developed an evaluation system and make the assessment of popularity of different computer science area.

In the experiment, we grouped over 200 GB on NYU Dumbo with Hadoop and Spark. As mention above, some ranking list such as the conference ranking in ERA in 2018 are mainly constrained in specific area or country. The problem of limited information can be solved by processing multiple datasets simultaneously and join same data (author, paper and venue) from different sources. Besides, our experiment also discarded outmoded and useless data. For example, we only concerned the ranking between 2008 and 2018 and discarded missing and inaccurate data, which we replace with number 0 or empty string. After that, we calculated our final score for different area in computer science and their changes in past ten years.

For the incorporation of venue information, the difficulty here is to find an official ranking list of venue. In that case, we have tried so many ranking list including the conference ranking proposed in 2010 by ERA (Excellence in Research for Australia), journal ranking list provided by Qualis in 2012, which is a Brazilian official system with the purpose of classifying scientific production and the conference ranking list provided by Microsoft Academic in 2014.

However, the conference ranking in ERA 2018 mainly constraint the venue in Australia, which is not suitable since it's only cover some of the venue information in our dataset. Ranking list provided by Qualis are mainly contains the journal, which maybe suitable for other field, but for computer science, lots of top conference are not in the list. For the conference ranking list provided by Microsoft Academic, although the list include the h-index that calculates the number of publications by author as well as the distribution of citations to the publications, and the field rating only calculates publications, citations as well as the impact of the scholar or journal within a specific field, it's evaluated in 2014, which may outdated in our case. However, the idea within it of taking h-index into consideration and evaluation is reasonable. Therefore, we used the h5-index and h5-median-index as our performance measure for venue.

Google Scholar is a freely accessible web search engine that indexes the full text or metadata of scholarly literature across an array of publishing formats and disciplines and the Google Scholar Metrics within it summarized recent citations of publications. Since there is no official download link and we cannot get access to all of the venue ranking list at a time, we simulate user request to crawl the venue information of each paper each time.

The experiment ranked popularity of 20 areas in computer science fields from 2008 to 2018, in which process the final area score have been generated. We plotted the ranking of area and corresponding year in graph, from which users can made some observations. For example, the machine learning appears to be increases since 2010, then it experienced a period of fluctuation but stayed in top 3. In 2016 the Information Theory was the hottest area in computer science field, but after that it was replaced with Machine Learning. Nowadays, Machine Learning is still the most popular topic among computer scientists. We have plenty of reasons to believe it will keep its leading position in next several years.

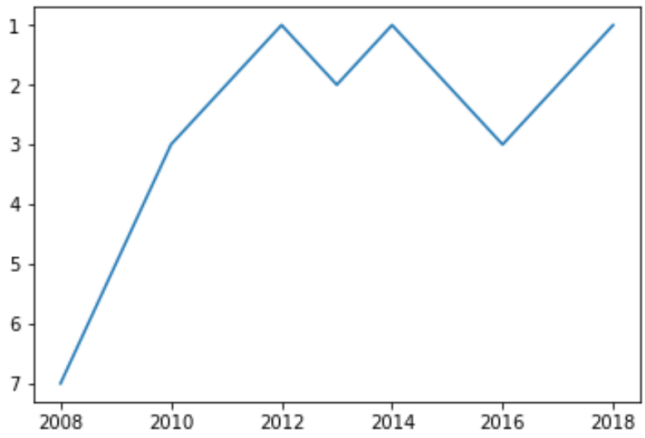

Fig. 2. Trend of area CS.LG (Machine Learning)

From our graphs, we can also make some other interesting observations: as the increases of popularity of Machine Learning field, some other fields related to Machine Learning closely including Data Structure and Algorithm, Discrete Mathematics, Computational Complexity, Computational Geometry and Databases are also getting a higher score. For example, after a long-term fluctuation, the popularity of Database are began to climbing in the ranking list since 2013, which is one of highlight year of ranking of Machine learning area. For machine learning researchers, the problem of the accessibility and processing of huge amount of data could not be ignored. Thus, it is reasonable to see the Database topic becoming more popular. Moreover, the popularity score of Game Theory and Social and Information Networks are also increasing though they are not so close to Machine Learning area.

Besides, our graphs showed that some other areas experienced a dramatic drop in recent years. Before 2013, the Information Theory was the most popular area in Computer Science field. However, it was replaced with Machine Leaning and dropped to 5th position in 2015. Although it came back to the leading position in ranking list in 2016, its ranking has still dropped in the next year. We can find another drop in graph of Distributed, Parallel and Cluster Computing since 2010. It started to recover from 2015. Recently, there are increasing number of machine learning scientists who are trying to incorporate distributed computing into the machine research. Our team believe there exists some relationships between these changes.

they appear concurrently particularly dynamic, with drastic changes and performance improvements witnessed each year, mainly due to the impact of artificial neural network (aka deep learning) models

VI. Future Work

In future, we plan to build the classifier (e.g. Logistic Regression, SVM, Random Forest) to predict the popularity score of each area. We also should incorporate the past reality data for comparison. For the training and evaluation, We plan

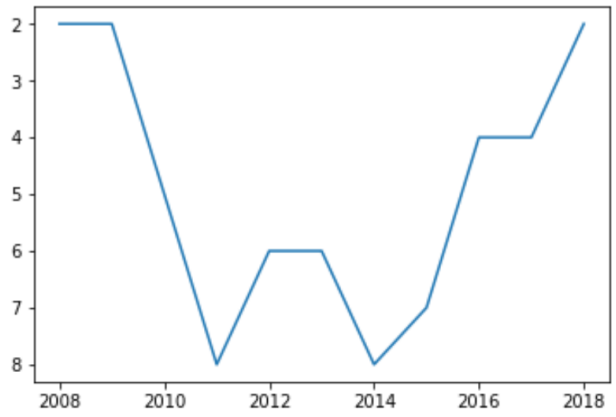

Fig. 3. Trend of area CS.DS (Data Structure and Algorithm)

Fig. 4. Trend of area CS.IT (Information Theory)

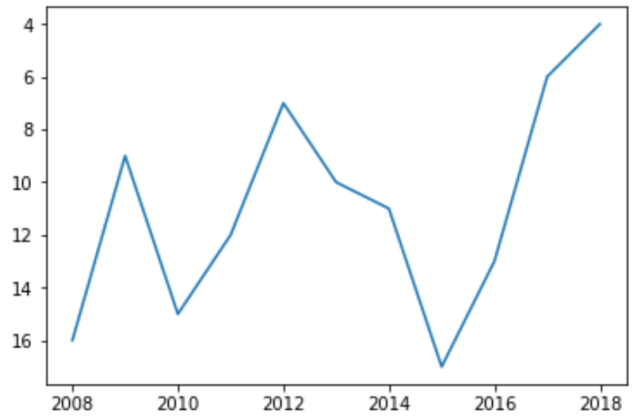

Fig. 5. Trend of area CS.DM (Discrete Mathematics)

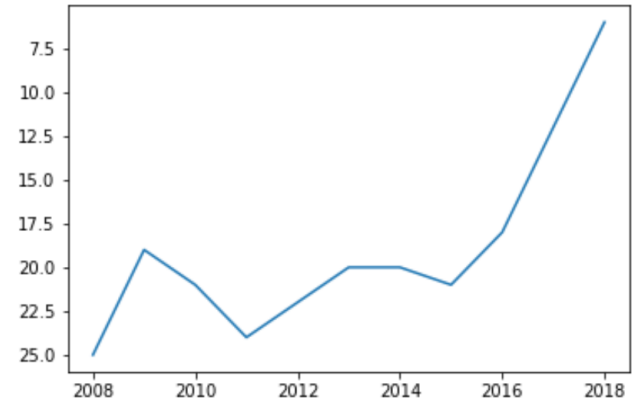

Fig. 6. Trend of area CS.CG (Computational Geometry)

to do cross validation to generate more reliable results. Also, the features analysis of classifiers will help us determine the confidence of the prediction result.

VII. Conclusion

We have developed a system to analyze the trend of hot areas in Computer Science by using the research information from arXiv and Open Academic Graph. By incorporating the paper information, the author information and the venue information within a specific area, our system are able to generate the popular score of the given area.

Acknowledgment

We would like to thank NYU High Performance Computing (HPC) for their support and for administering our Hadoop cluster, Dumbo. We would also like to thank Cloudera for providing the CDH Hadoop distribution through the NYU-Cloudera Academic Partnership.

References

[1] A. Duvvuru, S. Kamarthi, and S. Sultornsanee. Undercovering research trends: Network analysis of keywords in scholarly articles. In *2012 Ninth International Conference on Computer Science and Software Engineering (JCSSE)*, pages 265–270, May 2012.

[2] Gabriel Pui Cheong Fung, Jeffrey Xu Yu, and Wai Lam. News sensitive stock trend prediction. In Ming-Syan Chen, Philip S. Yu, and Bing Liu, editors, *Advances in Knowledge Discovery and Data Mining*, pages 481–493, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg.

[3] Qi He, Bi Chen, Jian Pei, Baojun Qiu, Prasenjit Mitra, and Lee Giles. Detecting topic evolution in scientific literature: How can citations help? In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 957–966, New York, NY, USA, 2009. ACM.

[4] Jose L. Hurtado, Ankur Agarwal, and Xingquan Zhu. Topic discovery and future trend forecasting for texts. *Journal of Big Data*, 3(1):7, Apr 2016.

[5] Jung-Hua Wang and Jia-Yann Leu. Stock market trend prediction using arima-based neural networks. In *Proceedings of International Conference on Neural Networks (ICNN'96)*, volume 4, pages 2160–2165 vol.4, June 1996.

[6] Dominique Lord and Bhagwant N. Persaud. Accident prediction models with and without trend: Application of the generalized estimating equations procedure. *Transportation Research Record*, 1717(1):102–108, 2000.

[7] Satyam Mukherjee, Daniel M. Romero, Ben Jones, and Brian Uzzi. The nearly universal link between the age of past knowledge and tomorrow's breakthroughs in science and technology: The hotspot. *Science Advances*, 3(4), 2017.

[8] Vinodkumar Prabhakaran, William L. Hamilton, Dan McFarland, and Dan Jurafsky. Predicting the rise and fall of scientific topics from trends in their rhetorical framing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1170–1180, Berlin, Germany, August 2016. Association for Computational Linguistics.

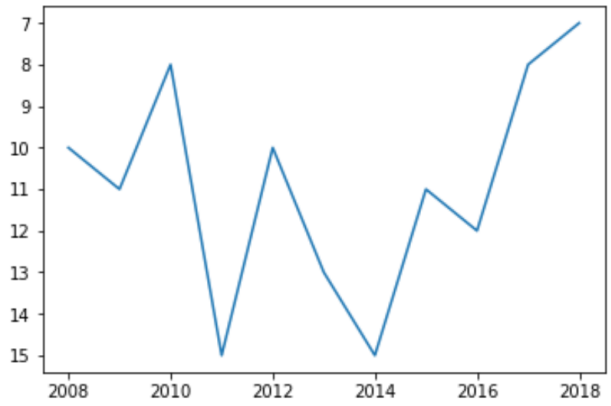

Fig. 7. Trend of area CS.CC (Computational Complexity)

Fig. 8. Trend of area CS.GT (Game Theory)

[9] Tieyun Qian, Qing Li, Bing Liu, Hui Xiong, Jaideep Srivastava, and Phillip C.Y. Sheu. Topic formation and development: a core-group evolving process. *World Wide Web*, 17(6):1343–1373, 1 2014.

[10] E. W. Saad, D. V. Prokhorov, and D. C. Wunsch. Comparative study of stock trend prediction using time delay, recurrent and probabilistic neural networks. *IEEE Transactions on Neural Networks*, 9(6):1456–1470, Nov 1998.

[11] José Carlos Oliveira Santos and Sergio Pazzini da Silva Matos. Analysing twitter and web queries for flu trend prediction. In *Theoretical Biology and Medical Modelling*, 2014.

[12] Naoki Shibata, Yuya Kajikawa, Yoshiyuki Takeda, and Katsumori Matsushima. Detecting emerging research fronts based on topological measures in citation networks of scientific publications. *Technovation*, 28(11):758 – 775, 2008.

[13] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, and Kuansan Wang. An overview of microsoft academic service (mas) and applications. In *WWW - World Wide Web Consortium (W3C)*, May 2015.

[14] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. Arnetminer: Extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 990–998, New York, NY, USA, 2008. ACM.

[15] Jinghao Zhao, Hao Wu, Fengyu Deng, Wentian Bao, Wencheng Tang, Luoyi Fu, and Xinbing Wang. Maximum value matters: Finding hot topics in scholarly fields. *CoRR*, abs/1710.06637, 2017.

Fig. 9. Trend of area CS.DC (Distributed Parallel, and Cluster Computing)

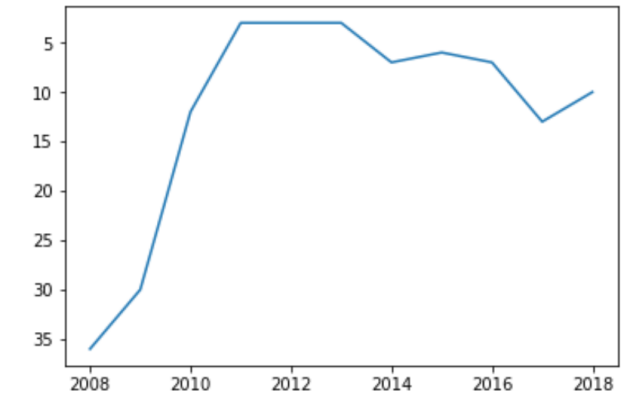

Fig. 10. Trend of area CS.SI (Social and Information Networks)

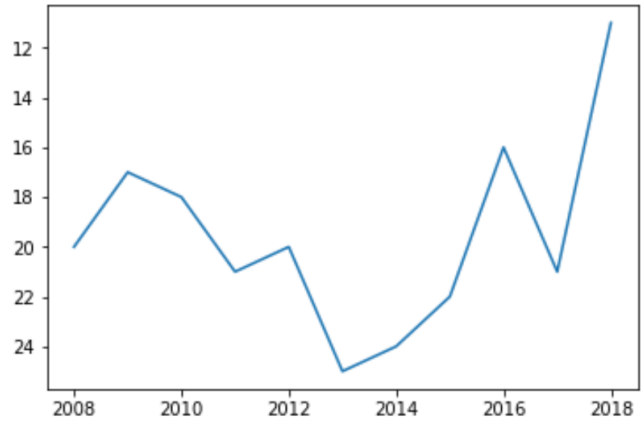

Fig. 11. Trend of area CS.DB (Databases)