Big Data Analytics Symposium - Summer 2019

Analytics Project: What's the trend of hot area in Computer Science field?

Team: Yijun Tian, Fan Xie, Rui Jin

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 area. This incurs the most interesting question - what's the trend of hot area 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 area 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 area reasonably. The experiment results show that our work performs well at showing the trend of hot area in computer science area.

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

Motivation

Who are the users of this analytic?

1. Computer Science field researchers
2. Newcomers who want to step into the computer field.
2. Venture capital

Who will benefit from this analytic?

1. Computer Science field researchers are able to study the changing trend of the whole industry.
2. Newcomers who want to step into the computer field can get a taste whats the most attractive area currently within the field
3. Venture capitalist can easily track hot topics and make appropriate assessments and investments.

Why is this analytic important?
The hot area point out the research interests of researchers, which is also a significant indicators to the strength of attention received from scientific communities.

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

Goodness

What steps were taken to assess the 'goodness' of the analytic?

We plot the ranking trend of the past hot areas and manually check top 20 hot areas. Results show:

1. The hot area in ranking list are really popular in real life.
2. The accuracy of internal rankings is reliable.

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

Data Sources

Name: arxiv

Description:

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. Almost all scientific papers are self-archived on the arXiv repository. Arrive API allows application developers to access all of the arXiv data, search and linking facilities with an easy-to-use programmatic interface.

Size of data: 10-100GB

Name: Open Academic Graph

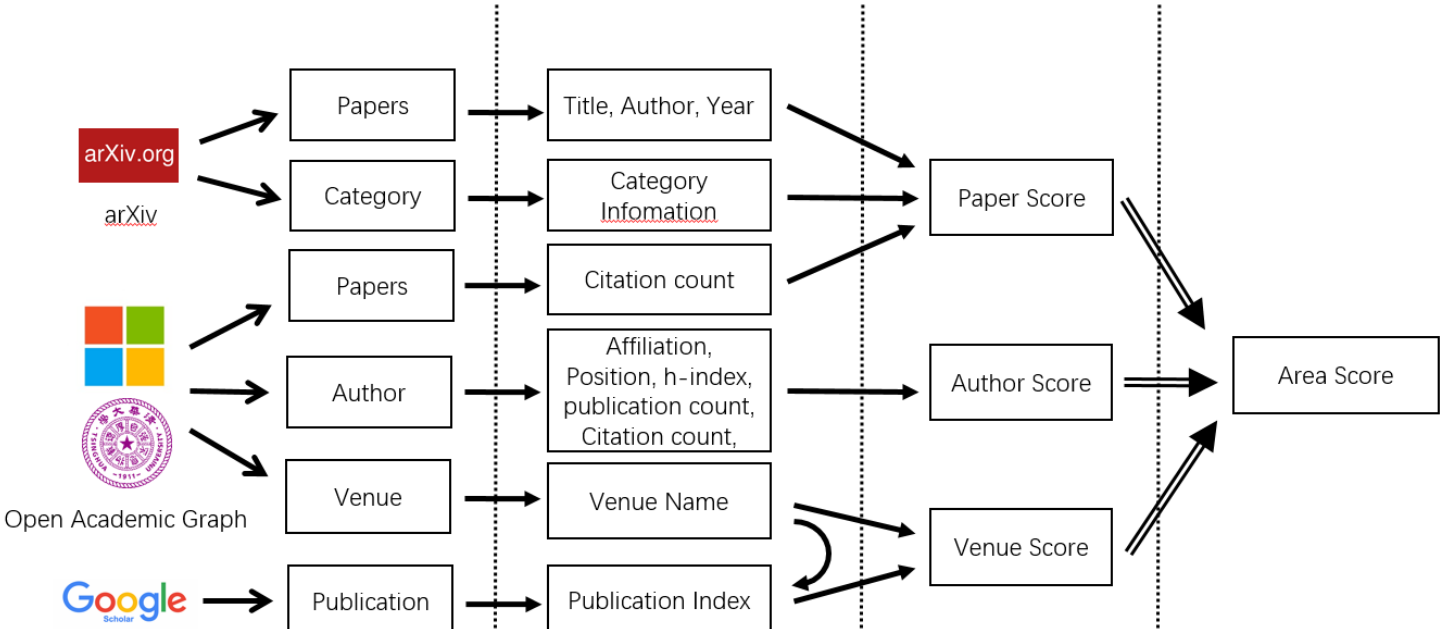
Description:

Open Academic Graph (OAG) [13], [14] contains 166,192,182 papers from MAG and 154,771,162 papers from AMiner (see below) and generated 64,639,608 linking (matching) relations between the two graphs. In OAG v2, author, venue and newer publication data and the corresponding matchings are available.

Size of data: 10-100GB

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

Design Diagram

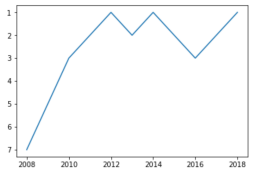


Platform(s) on which the analytic ran:

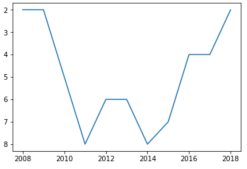
NYU HPC cluster

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

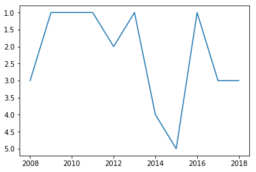
Results



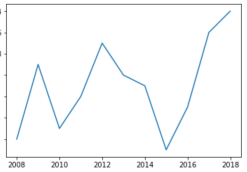
cs,lg: machine leanring



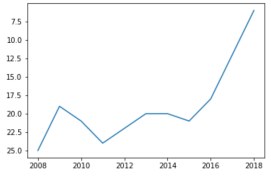
cs.ds: Data Structure And Algorithm



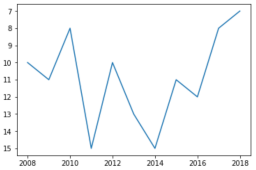
cs.it: Information Theory



cs.dm: Discrete Mathematics

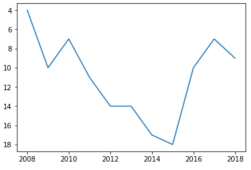


cs.cg: Computational Geometry

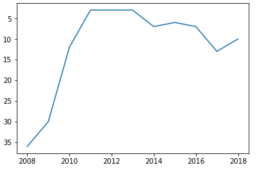


cs.cc: Computational Complexity

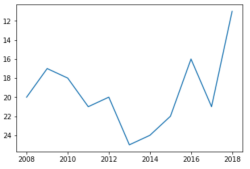
cs.gt: Game Theory



cs.dc: Distributed, Parallel, and Cluster Computing



cs.si: Social and Information Networks



cs.db: Databases

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

Obstacles

1. Cannot match some author score to paper score

2. Processing raw dataset (format of the data is not uniform)

Summary

We have developed a system to present the trend of hot area 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 popularity score of the given area.

Acknowledgements

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.

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

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. [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] Jose 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.

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

Thank you!