What's the next hot area in Computer Science field?

Yijun Tian *, Fan Xie *, Rui Jin *
New York University
Courant Institute of Mathematical Sciences
yt1506, fx349, rj1417@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. This incurs the most interesting question - what's the next 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 next hot area in the perspective of data is quite a crucial yet challenging problem. In this paper, we propose a novel technique that predict next hot area reasonably. The experiment results show that our work performs well at predicting next hot area in computer science area.**

***Index Terms*—Trend Prediction, Arxiv**

I. Introduction

Question like What are the next hot area in computer science research? appeared on Quora every year. People are enthusiastic to ask and answer 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. Most of the trend prediction are mainly based on some experts and there always existing some difference between each one. Besides, there is no consistent evaluation metric for the prediction is precise or not. Thus, predicting the next hot area in the perspective of data is natural, reliable and credible.

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 this case, newspapers, magazines and surveys publish trend ranking yearly. However,

topics in hot research area can attract people's 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 predict the hot areas in the future.

The main contributions of this paper are summarized as follows:

- we propose a novel schema that automatically predict next hot area in computer science field by analyzing and utilizing the past years data.
- We derive a new method to deal with the trend prediction 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 build a carefully designed user interface that visualize the result, which makes the exploration of the trend prediction 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 strength 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 prediction 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.

B. Trend prediction in academic

Satyam Mukherjee, Daniel M. Romero, Ben Jones and Brian Uzzi [7] parameterized the age distribution of works

* indicates equally contributed.

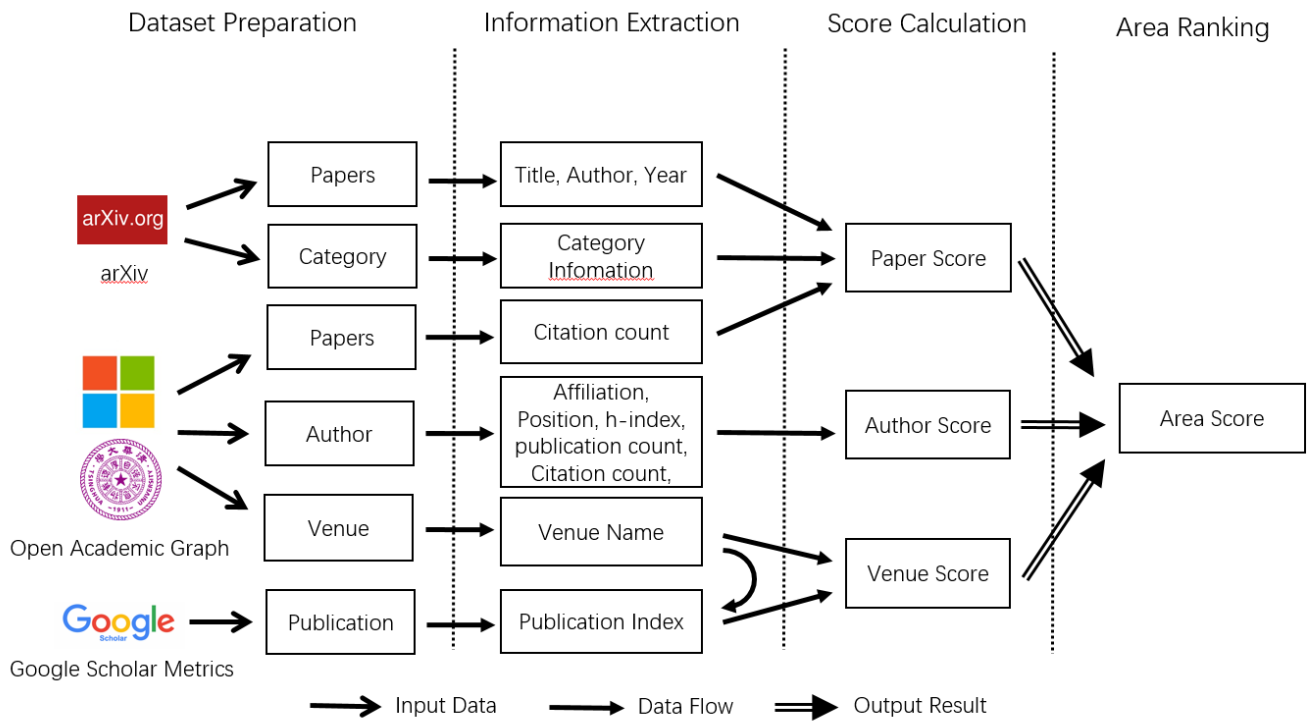

Fig. 1. Pipeline of prediction of next hot area in Computer Science field.

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. Application Design

We compare it with the past reality data. We can do cross validation on the dataset. Also, the features analysis will help us determine the confidence of the prediction result.

V. Datasets

A. *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. 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.

B. *Open Academic Graph*

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.

VI. Remediation

VII. Experiments

VIII. Conclusion

IX. Future Work

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.

[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.