Temporal Analysis of Scientific Literature to Find Grand Challenges and Saturated Problems

Kritika Agrawal and Vikram Pudi
kritika.agrawal@research.iiit.ac.in, vikram@iiit.ac.in

Data Sciences and Analytics Center, Kohli Center on Intelligent Systems IIIT,
Hyderabad, India

Abstract. As scientific communities grow and evolve, there is emergence of new techniques and decline of old ones. The tremendous amount of research publications available online aims to solve a lot of interesting problems. With time, some of the fields have been studied well and research problems solved to a great extent. However, there are few difficult research problems which are yet not solved completely and interests a lot of researchers. In this paper, we aim to find research fields which are saturated and research fields which need to be explored yet. We first extract research problems in a semi supervised manner using a proven bootstrap framework from scientific literature of the last fifty years. We show how a simple statistics based model on top of the research problems extracted can find the saturated fields and grand challenges in any domain of computer science.

Keywords: scientific data extraction · temporal analysis · unsupervised learning

1 Introduction and Related Work

A consistently thriving global research community has over decades produced a colossal amount of research papers that are published online, which makes it crucial to organize this huge bulk of information systematically so that upcoming researchers can navigate through efficiently and continue to push boundaries of scientific research. Such an organization over intellectual information will not only boost the rate of further research work but also augment researchers with a better holistic view of development in research and the directions in which it is evolving into. One of the first elementary steps we take as researchers is to figure out which problems to focus on solving, and structured analysis on present research status will help researchers identify critical problems and also give insight about how they developed across time. Due to this it will be easier to realize if particular problem has got no recent improvement in the recent past and has moved into a thriving application and so on. Analysis is the foundation to organization of cumulative knowledge garnered by the research community in decades, and this paper deals with this first step in direction.

[1] first proposed a task that defines scientific terms for 474 abstracts from the ACL anthology [2] into three aspects: domain, technique, and focus. They applied template-based bootstrapping on title and abstract of articles to tackle the problem. They used handcrafted dependency based features. Based on this study, [3] improved the performance by introducing hand- designed features to the bootstrapping framework. They both tried to study the influence of different scientific communities over the period of time. However, their work was limited to the computational linguistics field. We propose a method for temporal analysis of scientific literature of complete computer science domain.

A recent challenge on Scientific Information Extraction (ScienceIE) [4] provided a dataset consisting of 500 scientific paragraphs with keyphrase annotations for three categories: TASK, PROCESS, MATERIAL across three scientific domains, Computer Science, Material Science, and Physics. This invited many supervised and semi-supervised techniques in this field. Although all these techniques can help extract important concepts of a research paper in a particular domain, we need more general and scalable methods which can summarize the complete research community and help in time based analysis. For this we used a DBLP dataset which spans over fifty years and cover a wide variety of computer science fields.

As the first step of time based analysis, we aim to find saturated fields and grand challenges. We define saturated fields as those research problems which have been studied to a great extent and nothing much is left to achieve in them. On the other hand grand challenges are defined as those problems which have been tried to solve over a large period of time and are still worked upon extensively.

2 Definitions

Saturated Problems: Problems which were very actively studied in the yesteryears and are now solved to a great extent. Example, parts of speech tagging in NLP.

Grand Challenges: Problems which were defined in yesteryears and are still worked upon extensively. Example, machine translation in NLP. Research during the 1980s typically relied on translation through some variety of intermediary linguistic representation involving morphological, syntactic, and semantic analysis. In current times, research has focused on moving from domain specific systems to domain independent translation systems.

3 Approach

3.1 Identifying Aim and Method

Our approach is based on a proven method followed by [5] .Given a document, we classify its phrases as Aim or Method. This approach is built on the observation that the semantics of the sentence of a research article containing a phrase belonging to any of the concept type is similar across research papers. To capture this semantic similarity, we use k nearest neighbour classifier on top of state-of-the-art [6] domain based word embeddings. We start by extracting features from a small set of annotated examples and used bootstrapping framework [7] for extracting new features from unlabeled dataset. Finally, after some iterations, we have a set of phrases classified as Aim or Method for each research paper present in the dataset.

Merging of phrases which mean the same: We group the papers according to the conference in which they were published. Then ∀ papers in the same group, we cluster their extracted phrases by running DBSCAN [8] over vector space representations of these phrases. The clusters are created based on lexical similarity which is captured by cosine distance between phrase embeddings. [5] A cluster i belonging to conference c_1 and a cluster j belonging to conference c_2 are merged if they have any common phrase. Finally we get clusters such that phrases in each cluster have the same meaning.

3.2 Time based Analysis models

From the first step, we have research problems which have been studied as "AIM" for the last fifty years. We also have techniques"METHOD" used to solve these problems over these years. We first extract data for each research field, p, and find the number of times paper published on them for each of the years in the range 1971 to 2013.

Finding Saturated Problems:

- Count vs year plot for such problems should show a steep decline in the current years.
- Based on exploratory data analysis we came up with the following rules for finding saturated problems from the data collected above
- We list a problem p as a saturated problem if:
  - T_1 is the first year when the problem appeared in the literature. T_2 is the last time when the problem appeared in the literature.
  - Count of p appearing as aim in T_2 should be less than the count of p appearing as aim in T_1
  - Peak of count vs year plot should have occured much before 2013.
  - Suppose problem p1 has peak at time t_1 and problem p_2 has peak at time t_2. P_1 is a better candidate for saturated problem than p_2 if the difference between T_2 of p_1 and t_1 is more than the difference between T_2 of p_2 and t_2.

Finding Grand Problems:

- Count vs year plot for such problems should start from yester years and be consistent over the time. Peaks should be current years as well as yester years.
- Based on exploratory data analysis we came up with the following rules for finding grand challenges from the data collected above
  - We list a problem p as a grand challenge if:
    - T_1 is the first year when the problem appeared in the literature. T_2 is the last time when the problem appeared in the literature.
    - T_1 for problem p to be classified as a grand challenge should be before 2000 and T_2 after 2010. Time span between T_1 and T_2 should be more than 10 years.
    - Count of p appearing as an aim in T_2 should be more than some threshold. This is to rule out the edge cases where there is occurrence of few counts in current years.
  - We rank these problems based on the following formula:
    - To capture the fact that more the span of the problem over the years, more likely it is a grand challenge; we propose rank to be directly proportional to the number of years it spans to.
    - To capture the fact the count needs to be consistent over the years; we propose rank to be inversely proportional to $\sum_{i=1}^{n}(count[i] - count[i-1])$ where i iterates over all the years in which a problem p occurs.

$$Rank(p) \propto \frac{n}{\sum_{i=1}^{n}(count[i] - count[i-1])} \tag{1}$$

Where i iterates over all the years in which problem p occurs,starting from the second entry and n is the total number of years.

4 Experiments and Results

4.1 Dataset

All experiments were done on DBLP citation network version 7. We chose DBLP dataset to get a wide variety of research papers from different domains over a large time period. It has 2,244,021 papers and 4,354,534 citation relationships. After pruning out some papers and data cleaning we came up with 332,793 papers having 1,508,560 citation links. These papers range from 1936 to 2013. However for the period 1936- 1971, the number of papers available were relatively very less for time based analysis. So we pruned the data further and worked on papers from 1971 to 2013.

4.2 Finding Grand challenges and Saturated Problems:

We got a total of 555,383 problems in the first step. Out of these, our algorithm classified 599 as saturated problems and 1052 as grand challenges. To analyse the results, we extracted top 100 problems in both the categories. We represent our results as word clouds [9] where the font and color of each word is proportional to rank of that problem as extracted by our algorithm.

Grand Challenges	**Saturated Problems**
speech recognition	disk arrays
computer vision	schema integration
kolmogorov complexity	abductive reasoning
real-time applications	reconfigurable mesh
human-computer interaction	loop transformations
query language	non-monotonic reasoning
automatic parallelization	claw-free graphs
stereo vision	facility location problem
java	one-way function
xml	robot learning

Table 1. Top 10 Grand Challenges and Saturated Problems.

- **Discussion of Results:**
  1. Speech recognition has a rich history that precedes Internet era. In 1952, three bell lab researchers made "Audrey" which recognized formats in power spectrum of each word. Investment in research in this area amplified during 1970s with DARPA marking funding for understanding speech. IEEE speech groups were setup. In 1990s CMU led research funded Sphinx system which dominated DARPA 1992 evaluation. In 2005 Siri came into life under Apple. From 2012 there was a major breakthrough in research and HMM models which were industry standard till then were replaced by DNN. In 2014 end-to-end speech training was new paradigm that caught winds within DNN. In 2016 CMU and Google collectively introduced idea of "Attention" in training. In past three years there has been work on language agnostic ASR and more notable improvements kept on pressing. With importance of digital assistance, industry support has further expedited constant improvements every month over month till date. Clearly its a field with surreal active development and its not a surprise that our Model has correctly predicted this model as a Grand Challenge.
  2. Human Computer interaction is defined as a discipline concerned with the design and evolution of interactive computing systems for human use. HCI surfaced in the 1980s with the advent of personal computing, just as machines such as the Apple Macintosh, IBM PC 5150 started turning up in homes and offices. HCI soon became the subject of intense academic investigation. Initially, HCI researchers focused on how easy computers are to learn and use which has now also included to support the vision of personalized, adaptive, responsive, and proactive services,

adaptation and personalization methods and techniques that will need to consider how to incorporate AI and big data [10].

3. In algorithmic information theory, the Kolmogorov complexity of an object, such as a piece of text, is the length of a shortest computer program that produces the object as output. Research on this started in 1970s and is still going on.

4. The exact solution of facility location problem is known to be hard. And there are many approximation algorithms. No new research have been done on this problem. So clearly it is a saturated problem.

5. A **one-way function** is easy to compute on every input, but hard to invert. Although, The existence of true one-way functions is an open conjecture. In practice many functions such as those based on discrete Log are assumed to be work well since no polynomial time algorithm is known to invert them.

6. **Loop optimization** is the process of increasing execution speed and reducing overhead of loops. This problem is fairly solved and many modern compilers already use loop optimization techniques like Fission, Fusion, Inversion, Parallelisation etc.

Fig. 1. Word Cloud for Grand Challenges

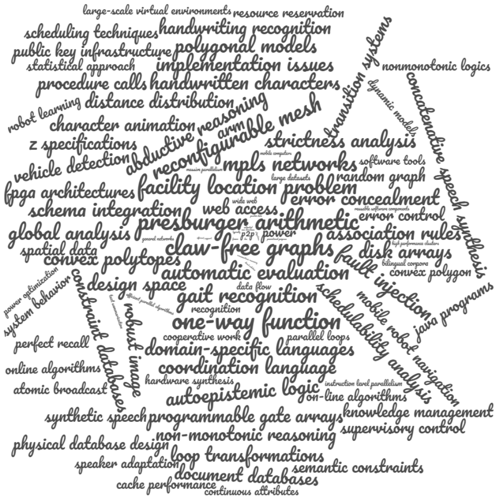

Fig. 2. Word Cloud for Saturated Problems

5 Conclusions and Next Steps

In this paper, we show the temporal analysis of scientific literature by extracting saturated problems and grand challenges. We propose this as the first step towards time based analysis. We plan to further do time based analysis by finding transition time for problems where transition time is defined as the time period where a problem starts occurring as method instead of aim.

References

1. Sonal Gupta and Christopher Manning. Analyzing the dynamics of research by extracting key aspects of scientific papers. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1–9, Chiang Mai, Thailand, November 2011. Asian Federation of Natural Language Processing.
2. Amjad Abu Jbara and Dragomir R. Radev. The acl anthology network corpus as a resource for nlp-based bibliometrics. 2013.
3. Chen-Tse Tsai, Gourab Kundu, and Dan Roth. Concept-based analysis of scientific literature. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, page 1733–1738, New York, NY, USA, 2013. Association for Computing Machinery.
4. Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555, Vancouver, Canada, August 2017. Association for Computational Linguistics.

5. Kritika Agrawal, Aakash Mittal, and Vikram Pudi. Scalable, semi-supervised extraction of structured information from scientific literature. In *Proceedings of the Workshop on Extracting Structured Knowledge from Scientific Publications*, pages 11–20, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
6. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.
7. Sonal Gupta and Christopher Manning. Improved pattern learning for bootstrapped entity extraction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 98–108, Ann Arbor, Michigan, June 2014. Association for Computational Linguistics.
8. Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.
9. F. Heimerl, S. Lohmann, S. Lange, and T. Ertl. Word cloud explorer: Text analytics based on word clouds. In *2014 47th Hawaii International Conference on System Sciences*, pages 1833–1842, 2014.
10. Chairs Constantine Stephanidis, Gavriel Salvendy, Members of the Group Margherita Antona, Jessie Y. C. Chen, Jianming Dong, Vincent G. Duffy, Xiaowen Fang, Cali Fidopiastis, Gino Fragomeni, Limin Paul Fu, Yinni Guo, Don Harris, Andri Ioannou, Kyeong ah (Kate) Jeong, Shin'ichi Konomi, Heidi Krömker, Masaaki Kurosu, James R. Lewis, Aaron Marcus, Gabriele Meiselwitz, Abbas Moallem, Hirohiko Mori, Fiona Fui-Hoon Nah, Stavroula Ntoa, Pei-Luen Patrick Rau, Dylan Schmorrow, Keng Siau, Norbert Streitz, Wentao Wang, Sakae Yamamoto, Panayiotis Zaphiris, and Jia Zhou. Seven HCI grand challenges. *International Journal of Human–Computer Interaction*, 35(14):1229–1269, 2019.