The Next Big Thing In Science

Adrian Mladenic Grobelnik
Artificial Intelligence Laboratory
Jozef Stefan Institute
Ljubljana Slovenia
adrian.m.grobelnik@ijs.si

Dunja Mladenic
Artificial Intelligence Laboratory
Jozef Stefan Institute
Ljubljana Slovenia
dunja.mladenic@ijs.si

Marko Grobelnik
Artificial Intelligence Laboratory
Jozef Stefan Institute
Ljubljana Slovenia
marko.grobelnik@ijs.si

ABSTRACT

This paper presents an approach to predicting the future development of scientific research based on scientific publications from the past two centuries. We have applied machine learning methods on the Microsoft Academic Graph dataset of scientific publications. Our experimental results show that the best performance is obtained for a noticeable increase of the topic frequency in the last 5 years compared to the previous 10 years. In this case, our model achieves precision of 74.3, recall of 71.7 and F1 of 73.0. Some topics that our model identified as promising are: *proton proton collisions, higgs boson, quark, hadron, mobile augmented reality, variable quantum, molecular dynamics simulations, hadronic final states, search for dark matter*.

CCS CONCEPTS

•CCS Information systems Information retrieval Document representation Content analysis and feature selection

KEYWORDS

Science analysis, machine learning, data representation

1 Introduction

With the ever-increasing pace of scientific developments, it is becoming difficult to keep track of current scientific research topics, let alone predict the promising lines for future research. As the quality and quantity of digitized scientific publications is growing, it has enabled modelling the development of scientific publications over time with greater accuracy and efficiency. In our research we explore how a simple Perceptron algorithm performs, given a considerable amount of data.

Our research hypothesis is that scientific topics that will be important in the future, already exist in today's scientific articles. To identify them, we applied machine learning methods on a large database of publications, namely the Microsoft Academic Graph [1]. We have defined a machine learning problem, such that the model predicts early indicators suggesting which scientific topics in today's literature will likely become important in the future.

In related work, researchers have addressed a similar problem also on a part of the Microsoft academic database of publications. They used a binary classifier to predict future developments in science. However, their research was on "Finding rising stars in academia early in their careers" [6]. Their representation comprises of authors' personal and social features. The research presented in [7] focuses on predicting emerging topics based on citation and co-citation data using clustering methods. The topics are classified to understand the motive forces behind their emergence ("scientific discovery, technological innovation, or exogenous events"). Emerging topics were also addressed in [8] where keywords from MeSH terms of PubMed database are filtered based on their increment rate of appearance in life science publications. In our research, we automatically generate frequent NGrams from the paper titles and use them to construct a machine learning model for predicting which topics will become popular in the future.

The main contributions of this paper are the proposed problem definition, data representation and the identified topics which are promising as the next big thing in science. The rest of the paper is structured as follows. Section 2 describes the data, Section 3 describes the problem, Section 4 presents the experimental results and Section 5 provides discussion.

2 Data Description

One could say the main element of science is an idea, invention or finding which occurs at the beginning of a scientific process. What follows is a period of scientific investigation, testing the idea in different contexts, proving the invention is useful or applying the findings in different scenarios. If proven to be valuable, new products or research is developed based on it. In our research, we rely on the fact that scientists are typically strict and consistent with naming conventions, enabling us to track the evolution of particular scientific topics through time.

In our research we have used the titles of scientific articles to identify when a scientific topic first appears, how frequently it appears through time, and when it stops being used. There are many databases of scientific articles in the world, but only some are open and available for research. Today, the biggest open database of scientific articles is known as the "Microsoft Academic Graph" which was released for research use in 2016. The database size is 104 Gigabytes, and it includes references to 125 million scientific articles from the year 1800 to 2015 from all areas of science. Each scientific article in the database is described by its: title, authors, their institutions, the journal or conference where it was published and the year of publication. The data is available from: https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/ .

.

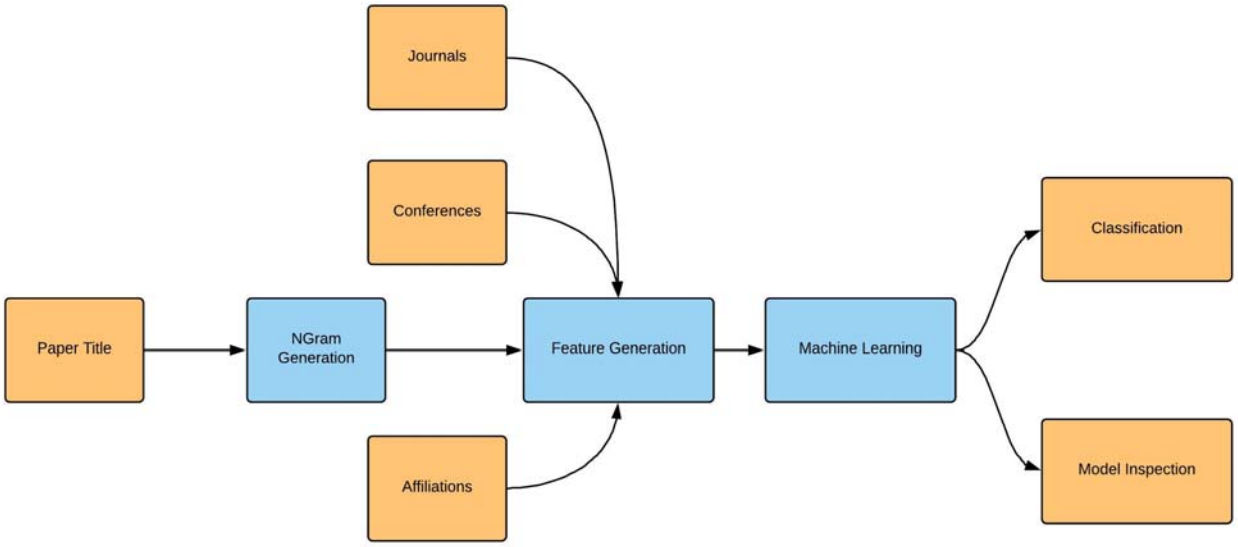

Figure 1: Architecture of the system including NGram extraction, feature generation and machine learning. The generated model is used for classification to predict the popular topics as well as to identify the most important features.

From the 125 million article titles, we extracted 2.5 million candidate topics, each corresponding to a phrase consisting of 1 to 5 consecutive words (also referred to as NGrams). The NGram must appear at least 100 times in the database of paper titles to be considered as a topic. Table 1 shows the distribution of NGrams

NGram	Total
1-Gram	300,000
2-Gram	1,000,000
3-Gram	800,000
4-Gram	300,000
5-Gram	100,000
All NGrams	2,500,000

Table 1: The number of NGrams generated from the publication titles

Figure 1 illustrates the process of feature generation and machine learning on the examples which represent the selected topics. The NGrams are generated from the paper titles, keeping only the frequent NGrams. For each frequent NGram, a feature vector is constructed using affiliations, conferences and journals of the papers in whose titles the NGram occurred.

For each topic, we find the longest span of years in which the topic appears in an article title at least once. Only topics which have the span of 15 years or longer are considered. This leaves us with about 1 million topics. Each topic is represented by a set of features describing the last 10 years before it became popular. The features include bag of affiliations, bag of journals and bag of conferences of the publications in which the topic occurred. For each topic we report the total frequency over the 10 years and the slope of a line through the (year, frequency) points.

For instance, “SVM” as a topic has occurred in papers published by authors affiliated with Oregon State University (slope 0.5), Max Planck Society (slope 3), University of Waterloo (slope -0.5). We can see that the popularity of the topic “SVM” in the Max Planck Society has increased within the observed 10 years.

Each topic is described by approximately 55,000 features (23,000 journals,1,300 conferences and 30,700 affiliations). Each topic is classified as either positive, if it became popular within the span of 15 years or as negative otherwise. Popularity is defined as a large difference in slopes of topic frequency in the 10 consecutive years compared to the following 5 consecutive years. We performed experiments varying the threshold (slope difference) from 1 to 5. A slope difference of 1 in our data results in 34% of examples being labeled as positive while a slope difference of 5 results in 20% of our examples being labeled as positive.

3 Problem Description and Algorithm

The problem we are solving is predicting early indicators suggesting which scientific topics are likely to become important in the future. The core task is to use the data from over 200 years of scientific discoveries from publications and to extract the early signs of a scientific topic becoming popular. Using machine learning algorithms, we have trained a statistical model to classify scientific topics into two categories: those which became important and those which did not. The model was trained on the data from

the year 1800 to 2015 to predict which topics will become relevant in the next 5 years from 2015.

For machine learning we used the Perceptron MaxMargin algorithm [2], an improved version of the perceptron algorithm. The improvement is in using two different margins, one for each class:

$$MinPosMargin = \frac{1}{\sqrt{BadPosExs}} \quad MinNegMargin = \frac{1}{\sqrt{BadNegExs}}$$

Where $BadPosExs$ and $BadNegExs$ are the numbers of misclassified positive and negative examples respectively in the previous epoch of training. In our experiments, we ran 3,000 epochs to build the model (meaning that we went through all the training examples 3,000 times). The learning rate was set to 0.02 in the case of no misclassifications in the previous epoch, and in the case of misclassifications, it was calculated as follows: $LearningRate = \frac{1}{\sqrt{BadPosExs+BadNegExs}}$

As we are training a linear model, by examining the model itself, we can see the weights assigned to the features. The higher the weight, the more important the feature for the positive class. This means that by examining the model, we can see which affiliations, journals and conferences contribute the most to a topic becoming popular in the future.

4 Experimental Results

We split the topics into a training (70%) and test set (30%), where the training set is used to train the model and testing set is used to test the model. The statistical model, trained with the MaxMargin Perceptron algorithm produced the following results on the testing data (see Table 2): Precision: 74.3 Recall: 71.7 F1: 73.0 for a slope difference of 1. This means the model correctly identifies 71.1% of the topics that became popular (recall) and 74.3% of the topics predicted to become popular really became popular (precision). As the slope difference increased the performance decreased, for instance, precision drops from 74.3 in slope difference 1 to 37.9 in slope difference 5. This is likely due to the increasing difficulty of the classification problem as the number of positive training examples decreases. The fact that the classification accuracy increases with the slope difference does not reflect improvement of the model's performance, as it is very close to the majority class (66% at slope difference 1, 80% at slope difference 5).

Slope Diff	Precision	Recall	F1	Accuracy
1	74.3125	71.6824	72.9737	63.1452
2	54.1432	60.3341	57.0712	60.1984
3	44.1246	46.7691	45.4084	69.2293
4	38.8584	47.1334	42.5978	76.6491
5	37.8595	45.1482	41.1838	82.9061

Table 2: Precision, recall, F1 and accuracy on test data for slope difference from 1 to 5.

Figure 2 shows the model's performance (estimated by a combination of precision and recall, F1) for 5 progressively stricter criteria of labelling topics as positive (slope difference 1-5).
We can see that the performance on the training and test set does not differ much on slope difference 1. As the slope difference increases, the performance on the test set drops relative to the performance on the training set.

Figure 2: Graph of model performance (measured by F1, the higher the better) for test and train data and 5 slope differences.

Looking at the resulting machine learning model we can see the following: if a scientific topic gets increasing attention from important research institutions (universities and research institutes), and is getting published by important journals and conferences within 10 years from its first appearance, then we can expect the increased use of the topic in scientific publications in the next 5 years.

In addition to the previous experiments, we have also built a perceptron model from scientific publications from 2006-2015.This model was used to predict future popular topics outside our dataset (5 years in the future from 2015). Looking at the results, one can notice several interesting topics predicted as promising. For instance: *proton proton collisions, higgs boson, quark, hadron, mobile augmented reality, variable quantum, molecular dynamics simulations, hadronic final states, search for dark matter.*

If we take a closer look at feature vectors of the promising topics during 2006-2015, we can notice for example that "*search for dark matter*" occurs in 56 papers with affiliation to *Purdue University* with a growing number of publications over the years (slope 4.14).

Another example is "*proton proton collisions*" which occurs in

- 610 papers with affiliation to the *Universite catholique de Louvain* with a growing number of publications over the years (slope 56.5).
- 8674 papers with affiliation to *CERN* with a growing number of publications over the years (slope 295.9).

Looking at the perceptron model trained on the data from 2006-2015, we can notice some of the most influential affiliations, conferences and journals are: *CERN, Journal of Proteomics & Bioinformatics, Industrial Research Limited, Circulation-*

cardiovascular Imaging, Molecular BioSystems, Metamaterials, Atw-international Journal for Nuclear Power, Data Science Journal, IEEE Geoscience and Remote Sensing Letters, Columbia college, Princeton university school of engineering and applied science.

5 Discussion

We analyzed 125 million articles from the "Microsoft Academic Graph" from over 200 years of scientific publications. In order to perform the experiments, we implemented the data preprocessing, feature generation and perceptron algorithm in C++. The resulting model was tested on a random 70/30 train/test split. The results show good performance, achieving F1 73.0%. The model predicts 71.7% of the scientific topics which became important in the history of science.

The possible direction for future work includes repeating the experiments on the new updated dataset, possibly considering the paper abstracts which have been made available in the dataset to be added to our feature set. It might also be beneficial to use the citation graph structure provided in the updated dataset. Another direction of future work would be applying the proposed approach to other similar datasets such as AMiner [3] or the Open Academic Graph [4, 5]. Yet another interesting direction of research would be to compare the performances of different machine learning algorithms and different data representations. Lastly, a more in-depth analysis of the topics predicted to become popular in the future would also be interesting.

We would also like to investigate ways to provide a publicly accessible online version of the system.

ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and DataBench European Union Horizon 2020 project under grant agreement H2020-ICT-780966.

REFERENCES

[1] *Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo - June(Paul) Hsu, and Kuansan Wang. (2015) An Overview of Microsoft Academic Service(MA) and Applications.In Proceedings of the 24th International Conference on World Wide Web(WWW '15 Companion). ACM, New York, NY, USA, 243-246.*

[2] Y. Li, H. Zaragoza, R. Herbrich, J. Shawe-Taylor, and J. Kandola. (2002) The perception algorithm with uneven margins. In *Proceedings of ICML 2002,* pages 379-386.

[3] AMiner (Accessed Sept 2019) https://aminer.org/

[4] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. (2008) ArnetMiner: Extraction and Mining of Academic Social Networks. In *Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (SIGKDD'2008). pp.990-998.

[5] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, and Kuansan Wang. (2015) An Overview of Microsoft Academic Service (MAS) and Applications. In Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion). ACM, New York, NY, USA, 243-246.

[6] Billah, Syed Masum & Gauch, Susan. (2015). Social Network Analysis for Predicting Emerging Researchers. 27-35. 10.5220/0005593500270035.

[7] Small, H., Boyack, K. W., and Klavans, R. (2014). Identifying emerging topics in science and technology. Research Policy, 48(8):1450–1467 – Elsevier.

[8] Ohniwa, R., Hibino, A., & Takeyasu, K. (2010). Trends in research foci in life science fields over the last 30 years monitored by emerging topics. *Scientometrics*, *85*(1), 111-127.