Towards predicting academic impact from mainstream news and weblogs: A heterogeneous graph based approach

Towards Predicting Academic Impact from Mainstream News and Weblogs: A Heterogeneous Graph Based Approach

Mohan Timilsina*, Brian Davis*, Mike Taylor†,Conor Hayes*
*Insight Centre for Data Analytics
National University of Ireland Galway
{mohan.timilsina, brian.davis, conor.hayes}@insight-centre.org
†Elsevier Informetrics
Oxford, United Kingdom
{mi.taylor}@elsevier.com

***Abstract*—The realization that scholarly publications are discussed and have influence on discourse outside scientific and academic domains has given rise to area of scientometrics called alternative metrics or "altmetrics". Furthermore, researchers in this field tend to focus primarily on measuring scientific activity on social media platforms such as Twitter, however these count-based metrics are vulnerable to gaming because they tend to lack concrete justification or reference to the primary source. In this collaboration with Elsevier, we extend the conventional citation graph to a heterogeneous graph of publications, scientists, venues, organizations and more authoritative media sources such as mainstream news and weblogs. Our approach consists of two parts: one is integrating the bibliometric data with the social data such as blogs, mainstream news. The other involves understanding how standard graph-based metrics can be used to predict the academic impact. Our result showed the computed graph-based metrics can reasonably predict the academic impact of early stage researchers.**

I. Introduction

The current trend of measuring impact of scholarly activity outside the scientific community is based on count of bookmarks, blog posts, views, tweets, likes, shares, etc. The activity in social media like Facebook or Twitter is quite benign, merely pointers to research rather than providing comments or discussion [1], [2]. In order to address this apparent weakness in social media, we chose to use lengthier document, for example, blogs and mainstream news references. Due to this, we proposed an approach to measure the impact of scientists in the non-scholarly literature as a means to measure their social impact and to predict academic impact.

II. Data Collection

Social media dataset: Our data is called a *Spinn3r*[1] data which is a crawl of the blogosphere for the time period of (Nov 2010-July 2011). We extracted only weblogs and mainstream news using Java Spinn3r API[2] and the collected data are stored in a MongoDb[3] database which stores the data as JSON[4] documents.

Bibliometric dataset: For the biblometric data, we used Scopus, one of the largest bibliographic database which contains the citations of peer-reviewed literature: scientific journals, books and conference proceedings. We used the Elsevier SCOPUS API[5] to extract metadata of publications such as citations, authors, publication venue and organizations.

Search of a Candidate Topic: We used Wikipedia to research prominent news events recorded in the time window of our social media index (Nov 2010-July 2011). This suggested one public health topic was particularly newsworthy: The emergence of virulent strain of *"Avian Influenza"*. We collected 259,149 JSON documents of total size 23 GB from Spinn3r dataset and 37,081 scientific publications from Scopus dataset.

III. Methodology

Construction of Graph Data Model: For this task, we used the conceptual model of graph data from the system architecture of Targeted Elsevier Project at *Insight Centre for Data Analytics*. The conceptual model consists of seven different types of node entities and five different types of relationships entities. The figure 1 shows the graph data model used for storing the data and for analysis:

Data Integration Process: We applied a custom text analysis pipeline based on the Natural Language Processing (NLP) framework **GATE**[6] over the contents of Spinn3r data which are blogs and mainstream news and identified **320** names of the scientist.

We used these names in a Scopus graph and manually verify checking their publications related to *"Avian Influenza"*. The identified **Scientist** are then connected to **Scientist** node of Scopus graph through **hasMention** relationships.

[1]http://spinn3r.com/
[2]http://www.programmableweb.com/api/spinn3r
[3]https://www.mongodb.org/
[4]http://www.json.org/
[5]http://dev.elsevier.com
[6]https://gate.ac.uk/

IEEE/ACM ASONAM 2016, August 18-21, 2016, San Francisco, CA, USA
978-1-5090-2846-7/16/$31.00

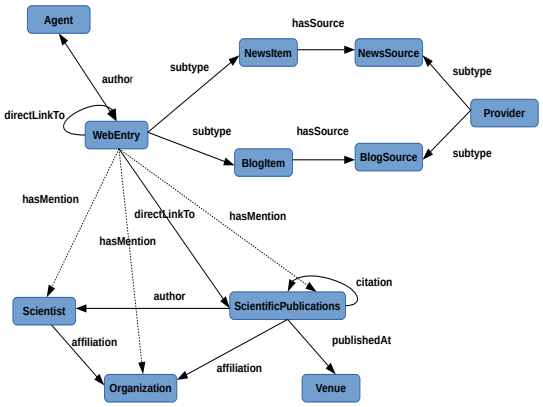

Fig. 1. *Conceptual Graph Data Model*

Maximal Directed Ego Network: We extracted the maximal directed ego network of scientist who are mentioned in a blogs or a news. The definition of maximal directed ego network is given as:

Definition 1: *Maximal Directed Ego Network*: A maximal directed ego network of a graph $G = (V, E)$ is an ego network of k hop away from the node s_i given by $G_i^k = (s_i \cup V_i^k, E_i)$ such that there is no vertex in $V \setminus V_i^k$ whose addition in G_i^k would preserve the property of a directed ego centered network.

Log Based Weight: This is the metric we proposed to weight the nodes in a maximal directed ego network. Log based weight is based on the information spreading ability of the nodes. If a scientist is mentioned in a maximal directed ego network then total influence of the scientist in that network is the cumulative sum of spreading ability of each nodes which is given by,

$$Scientist_(influence) = \sum_{i=1}^{N} log\Big[\frac{Indegree + 1}{Outdegree + 1} + 1\Big] \quad (1)$$

We applied this metrics and computed the scores for 320 scientist and compared them with their h-index. The result showed significant spearman correlation (ρ = 0.45 and p-value = 2.2e-16). Our next step is to predict the class of scientist.

Predicting the Class of a Scientist: We used the feature from *Maximal Directed Ego Network* namely depth of the graph, total number of nodes in the graph, cosine similarity between the citing and cited documents, number of mentions and the Log Based Weight. The outcome variable is the classes of the scientist as shown in the table I.

Quartile Distribution	Category
(0-25)%	Low Cited
(25-50)%	Moderately Cited
(50-75)%	Highly Cited
75% above	Very Highly Cited

TABLE I
Classification of Scientist with the Quartile Distribution of their h-index

We applied *Support Vector Machine algorithm (SVM)* in our datasets using 75 % as training set and 25 % as testing set.

Result: The result of the classification is shown in the figure 2:

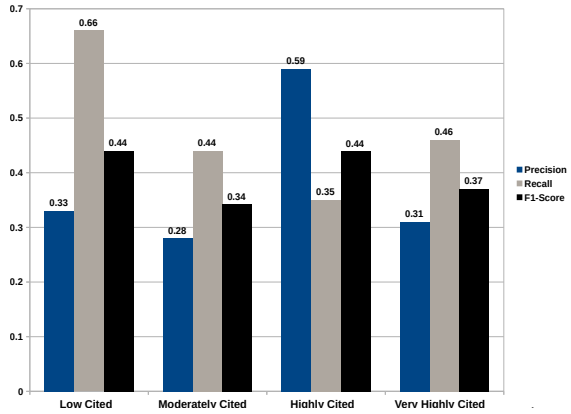

Fig. 2. *Result of the Classification*

In comparison with other class the precision of the model is higher for prediction of **Highly Cited** class which is 0.59 and lower for the Moderately Cited class as 0.28. Similarly, the precision for predicting **Low Cited** and **Very Highly Cited** class is 0.33 and 0.31 respectively. In correspondence with other class the recall of the model is higher for predicting **Low Cited** as 0.66 and lower for **Highly Cited** class as 0.35. Furthermore, recall for **Moderately Cited** is 0.44 and **Very Highly Cited** class is 0.46. In the context of balance between precision and recall we observed that the model has high F1 score of 0.44 for predicting **Low Cited** and **Highly Cited** class and low of 0.34 for **Moderately Cited** class. Similarly F1 score for predicting **Very Highly Cited** class is 0.37.

IV. Conclusion

The model struggled to correctly classify **Very Highly Cited** scientist. One reason for this may be because of the nature of the h-index which is the higher the h-index value, the harder it is to increase it. In such case including other academic factors of a scientist such as number of publications that gets citations, recognition of the work in the academic community, speaker in a scientific conference, seniority in the research field etc would have improved classification result. We would like to explore this in our future work.

Acknowledgment

This work has been funded by Scientific Foundation of Ireland (SFI/12/RC/2289) and the targeted project Elsevier. We appreciated Dr. Jonice Oliveira from Federal University of Rio de Jeneiro for creative feedback and support.

References

[1] M. Taylor, "The challenges of measuring social impact using altmetrics [internet]. research trends 2013 jun [cited 2014 feb 19]; 33: 11-15."
[2] D. Colquhoun and A. Plested, "Why you should ignore altmetrics and other bibliometric nightmares," 2014.