Identification of Key Opinion Leaders in Pharmaceutics Using Network Analysis

Sarthak Kulkarni[1], Pritam Bagad[1], Hrishikesh Joshi[1], Himanshu Randad[1] , Prof. Anuja Phapale[2]

[1]B.E., Information Technology Department, AISSMS Institute of Information technology, Pune, Maharashtra, India

[2]Assistant Professor at Information Technology Department, AISSMS Institute of Information technology, Pune, Maharashtra, India

ABSTRACT

The term Key Opinion Leader in marketing is not new. Key Opinion Leaders (KOLs) commonly known as thought leaders who play a crucial role in the life science industry. We through this project intend to implement the concept of identifying key opinion leaders using weighted Social Network Analysis (SNA). We intend to use European PubMed Central dataset for creating a weighted social Network of authors who have healthcare and medicine related publications and apply different centrality measures on it. In order to collect the data, we will be using one of the web scraping methods and predefined libraries like scrapy. After fetching and processing the data we intend to form a network of authors using python's NetworkX library. This network will then be subjected to various centrality measures which in turn will give prominent opinion leaders as the output.

Keywords : Network Analysis, Centrality algorithms, web scraping, network analysis, key opinion leaders, pharmaceutics

I. INTRODUCTION

The pharmaceutical industry plays a key role in the development of medical sciences. From discovering various drugs to their development, production and management pharmaceutical companies are responsible for all these processes. But developing drugs that can be used as medications for patients to cure them is not an easy task. It could take many years for studying, developing new medicines or drugs and their use in real life. As time plays important role in this process key opinion leader's knowledge can become a great asset for reducing time and cost, increasing efficiency as well. Key opinion leaders have knowledge and experience in their respective field of work and their research work can be helpful for pharmaceutical companies. As there can be a lot of researchers around the globe it can be difficult for the industry while hiring and selecting the right candidate. In this paper, we are proposing a

platform that will help the industry for selecting the right and eligible candidate. For this, we will be using data provided by European PubMed for gathering information about researches, their work in that field, publications they have done as well as citations. From this information and data, depending on citations we will be creating a network of researchers and will be analyzing this network by applying algorithms that are used in graph theory for network analysis like degree centrality, closeness centrality, betweenness centrality and eigenvector centrality [4].

II. LITERATURE SURVEY

Hengmin Zhou, Daniel Zeng and Changli Zhang [1] have used the opinion rank algorithm to find leaders. To evaluate the performance of the proposed Opinion Rank method, they have compared it with Page Rank, Random Sampling, Activity based rank and registering time. Out of the above four, Activity based rank method outperformed the Opinion Rank Method so according to it they add some content of Activity based rank method in their Opinion Rank Method and do useful modifications in the opinion rank method and performance becomes better of opinion rank method. This algorithm works but for only small datasets, more work needs to be done for larger datasets.

Huanhuan Liu, Xiaoqing Yu, Jing Lu [2] have used synthesis centrality algorithms to identify top N opinion leaders on a local social network. After comparing accuracy synthesis centrality algorithm outperformed page rank and HITS algorithm. The accuracy of synthesis centrality increases with the number of opinion leaders increasing.
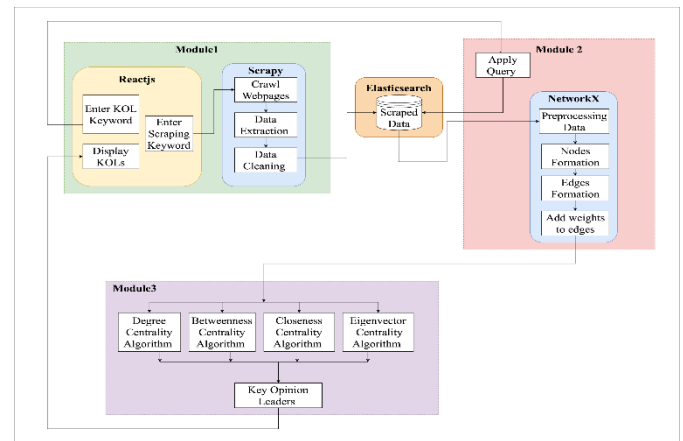
III. IMPLEMENTATION OF MODULES

Figure 1. Block Diagram

A. Frontend Technologies

Technologies which are used while building the client in our project are as follows:

1. Reactjs: It is an open-source JavaScript library used in the frontend for building reusable UI components as well as interfaces. It encourages the creation of reusable UI components, which present data that changes over time [7].
2. Redux: Redux is a predictable state container designed to help you write JavaScript apps that behave onsistently across client, server, and native environments and are easy to test [8].
3. D3: D3 is a JavaScript library for visualizing data using web standards. D3 helps you bring data to life using SVG, Canvas and HTML. D3 combines powerful visualization and interaction techniques with a data-driven approach to DOM manipulation, giving you the full capabilities of modern browsers and the freedom to design the right visual interface for your data [9].
4. *Bootstrap*: Bootstrap is a free and open-source CSS framework directed at responsive, mobile-first front-end web development. It contains CSS- and (optionally) JavaScript-based design templates for

typography, forms, buttons, navigation, and other interface components [10].

B. Backend Technologies

The following are technologies used to build the server:

1. Django: Django is a high-level Python web framework that encourages rapid development and clean, pragmatic design. Django makes it easier to build better web apps quickly and with less code [11].

2. Scrapy: Scrapy is a fast, open-source web crawling framework written in Python, used to extract the data from the web page with the help of selectors based on XPath [12].

3. NetworkX: It is a package used while creating and manipulating various complicated networks and also helpful for understanding and studying the dimensions and dynamics of these networks. Various data Structures for graphs are included in NetworkX. By using NetworkX we can construct random graphs, can find subgraphs, can convert graphs from one form to another. When operation on large real-world graphs is required NetworkX is really useful as it is reasonably efficient and scalable [5].

4. Elastic: Elasticsearch is a search engine based on the Lucene library. It provides a distributed, multitenant-capable full-text search engine with an HTTP web interface and schema-free JSON documents [13].

MODULE 1

Module 1 consists of a Web App where a user will interact with the system. If a user enters a keyword to find KOLs and it is not present in the database then research papers for that keyword will be scraped from websites. The data will be cleaned and then saved in the database. When research paper data is available then entered keyword is passed to module 2 for network formation.
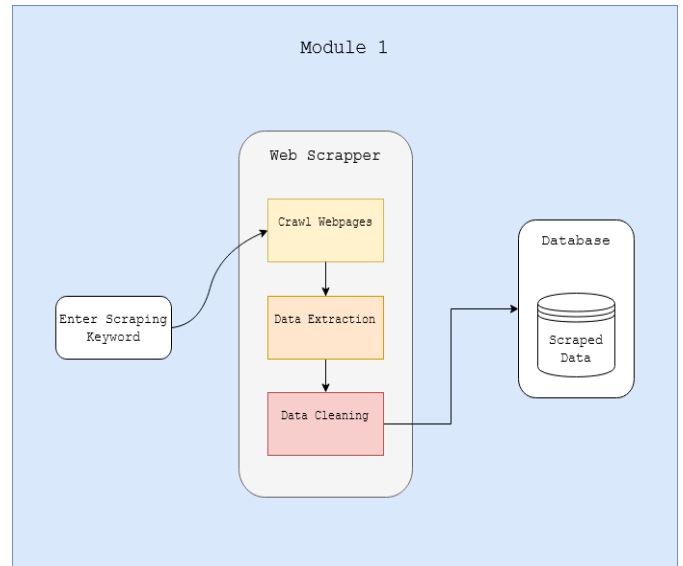

Figure 2.

MODULE 2

The data fetched from the database consists of research papers related to the keyword entered. We will use the title, authors and citations of each paper in the retrieved data. Now each author is represented using a node in the network. Every edge represents at least one common publication between the authors. So, the weight of an edge between authors is equal to the number of citations for that paper. If authors have contributed to more papers together then edge weight will be summation of citations of those papers.
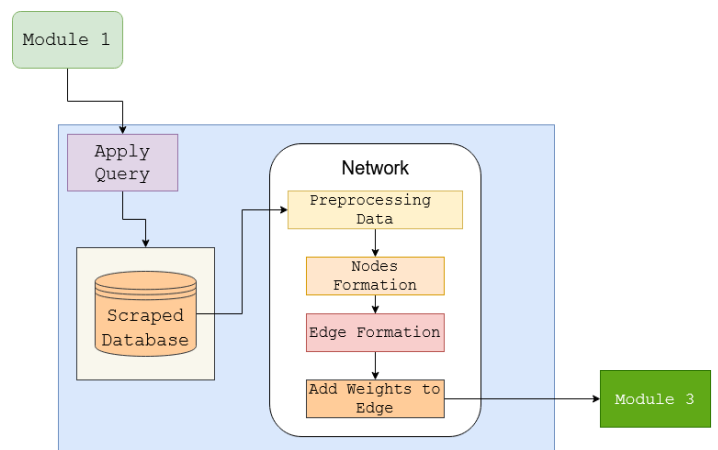

Figure 3

MODULE 3

Various algorithms which are used for analyzing networks are described in this module. The algorithms include [3]:

1. Degree centrality: It is a network analysis algorithm in graph theory used for determining how many nodes a specific node is directly connected in graph. More the degree of node, the central is the node in graph.

2. Closeness centrality: In graph theory, this algorithm is used for determining how close a node is with respect to other nodes. We can calculate the closeness centrality by

calculating average of shortest path available between given node and remaining every other node. The lesser the value of average, closer and more central the node is in graph.

3. Betweenness centrality: It helps in determining how much a given node in graph is present in-between the other nodes. It is determined by calculating how much the given node is present in the shortest path between other two nodes as compared to all shortest path existing between two nodes. Higher the value of betweenness centrality of specific node greater the importance of node for passing information in the network.

4. Eigenvector Centrality: In network, not all nodes have same relevance as well as importance. Some nodes are more important that others and these nodes are also considered as influential nodes in network. This algorithm works on principle a node is important and relevant compared to other nodes if it is linked to by other influential and important nodes in network.
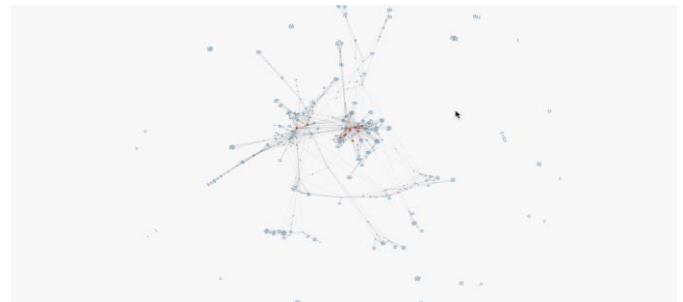
IV. OUTPUT

Figure 4. Complete Network

Once the data is scrapped it is cleaned and stored in the database.Redux Library integrated with ReactJS is used for app state management .Clients sends POST HTTPS and GETS requests to the server and get a response in JSON format using restful API, .Then formed network is exported from NetworkX library as a JSON document and is stored in database. D3 visualization library is used to display the formed network on WebApp.

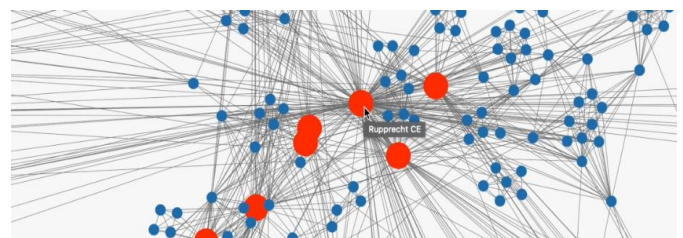

Figure 5. Database for keyword 'rabies'

Figure 6. Network with author Rupprecht CE

Figure 7. Network with author SHIRAKI

Here, we have scraped 249 research papers for keyword Rabies. Consider one author named Rupprecht CE. He has contributed to many research papers so in the network, he has many edges connected to other authors. Now consider author SHIRAKI H, he has only contributed to one research paper with one other author so in the network we can find just the node with one other node associated with one other node.

V. CONCLUSION

We have produced this report after studying the literature related to our project topic. After studying the literature, we have proposed a system for the identification of Key Opinion Leaders in Pharmaceutics using Network Analysis. This system is divided into three modules. The first module contains data scraping and collection. The second module is network formation using the Networkx library where each node represents the author and edges represent co-authored publications. The last module is to identify Key Opinion Leaders by applying centrality algorithms (degree, betweenness, closeness, eigenvector) on the network formed in the second module.

V. REFERENCES

[1]. Hengmin Zhou, Daniel Zeng and Changli Zhang, "Finding leaders from opinion networks," 2009 IEEE International Conference on Intelligence and Security Informatics, Richardson, TX, USA, 2009, pp. 266-268, doi: 10.1109/ISI.2009.5137323.

[2]. Yu, Xiaoqing & Lu, Jing & Liu, Huanhuan. (2013). Identifying top-N opinion leaders on local social network. IET Conference Publications. 2013. 268-271.10.1049/cp.2013.1970.

[3]. Identifying Key opinion leaders using social network analysis. Cognizant System Private Limited's White Paper in June 2015.

[4]. Network Analysis algorithms: Network analysis algorithm, Closeness centrality.

[5]. Python NetworkX - https://networkx.github.io/

[6]. Europe PubMed Central - https://europepmc.org/

[7]. ReactJS - Overview - Tutorialspoint

[8]. Redux - https://www.tutorialspoint.com/redux/index.htm

[9]. D3 JS - https://en.wikipedia.org/wiki/D3.js

[10]. Bootstrap - Bootstrap (front-end framework) - Wikipedia

[11]. Django-Django - Basics – Tutorialspoint

[12]. Scrapy - Scrapy Tutorial - Tutorialspoint

[13]. Elastic - Elasticsearch - Wikipedia

Cite this article as :