Hybrid Knowledge Mining Ecosystem

Robin Ghosh
Jackson State University
1400 John R Lynch St
Jackson, MS-39217
(+1) 870-329-2255
robin.ghosh@students.jsums.edu

Nicholas Gonzalez
Jackson State University
1400 John R Lynch St
Jackson, MS-39217
(+1) 314-562-0845
nicholas.gonzalez_del_ca@students.jsums.edu

Ciji Ramos
Jackson State University
1400 John R Lynch St
Jackson, MS-39217
(+1) 505-917-3744
ciji.a.ramos@students.jsums.edu

Jessie Walker
Jackson State University
1400 John R Lynch St
Jackson, MS-39217
(+1) 601-979-2059
jessie.j.walker@jsums.edu

Amanuel Gebre
Jackson State University
1400 John R Lynch St
Jackson, MS-39217
(+1) 336-605-8386
amanuel_engeda.gebre@students.jsums.edu

MohiuddinHasan
Jackson State University
1400 John R Lynch St
Jackson, MS-39217
(+1) 870-329-5796
md.m.hasan@students.jsums.edu

ABSTRACT

Scientific discovery today depends as never before upon ease of access to data, associated sophisticated tools and applications, to enable research and education. Researchers who once worked in local, isolated laboratories now collaborate routinely and on a global scale. Specialized instruments that were spread across multiple locations can now fit into a single lab connected via cyberinfrastructure resources and residing in big data. However, the sheer volume and heterogeneity of data bring a multitude of problems. Commodity scalable cluster-based platform enables diversity of researchers to efficiently extract, store, index, annotate, mine, curating, and search large scale diverse, heterogeneous open data source data sets such as Google Scholar [1]. This paper outlines the process of developing a hybrid commodity cluster, leveraging the commercial service IBM Bluemix via Watson Data Analytics.

MEDES '17, November 7–10, 2017, Bangkok, Thailand

ACM ISBN 978-1-4503-4895-9/17/11$15.00
https://doi.org/10.1145/3167020.3167024

CCS Concepts:

• **Information systems~Data mining** • *Information systems~Data analytics* • Computing methodologies~Artificial intelligence

Keywords

Big data, Data Mining, Clustering, Data Analytics and Visualization

1. INTRODUCTION

Collaboration in research and crowd-sourced learning has accelerated the quality and number of research as well as scientific publications in various disciplines all over the world. New scientific publications for new findings and research in various topics are being shared and stored at a rate that is growing exponentially. This has caused researchers to become overwhelmed with the sheer number of scientific documents shared and stored. One of the methods in a research attempting to estimate the amount of research publication circulating in the academic web estimates 114 million number of circulating documents written in English just up to 2014 [2]. This estimate is just a portion of what was out here during 2014 and it surely has increased much more since then. As a result, scientists need to keep up with all the new developments in their area by reviewing and analyzing scientific publications and reports in order to have well-rounded understanding of interdisciplinary works. With the ever-growing scientific

publications and documents, it is a daunting task for scientists and researchers to analyze and draw a conclusion. This shows that the need of Knowledge mining techniques and algorithms, as well as an infrastructure to address this issue, becomes ever more evident.

The knowledge mining of research papers and scientific publications is not a new idea but to access large amounts of research papers were controlled by a handful of companies having bespoke arrangements with publishers. The Open Access movement has recently largely contributed to decreasing the barriers to text mining of research papers [3]. In fact, there are few types of research regarding this area, which use Natural Language Processing (NLP) to semantically analyze and explore data as well as summarize publications. But, previous researches regarding knowledge mining of research papers do not fully address the logistics of on how to actually deploy a feasible cluster cloud infrastructure to store, analyze and draw insights by mining publications. The possibility to utilize and integrate readily available well-developed commercial cloud computing resources to analyze documents has also not been fully explored.

In this paper, a hybrid knowledge-mining ecosystem is proposed to address knowledge mining from scientific publications first, by pulling indexed publications specifically from Google Scholar as needed on private robust and highly scalable cloud infrastructure using Open Stack [4] on commodity hardware. Second, by using powerful commercial data analysis tools on IBM Watson Analytics such as Natural Language Understanding (NLU), Watson Discovery, Retrieve and Rank (R&R). The use of commodity hardware ensures the economic feasibility of the project using cheaper commodity hardware, which means the infrastructure, can easily be set up in university research labs. The system integrates Application Programming Interfaces (APIs) to access commercial cloud services, which provide services to analyze required research documents and also provide flexibility to use those reliable services that offer free or with little cost.

2. RELATED WORK

There were a lot of techniques of text mining research papers, such as: Literature-Based Discovery (LBD), summarization of research findings, question answering, semantic search from papers and others [3]. Things have been done before for knowledge extraction and modeling from scientific publications using NLP. Dr. Inventor project [5] identifies creative analogies across papers. The framework is distributed as a self-contained Java library, thus providing a convenient tool both to bootstrap more complex scientific publication analysis experiments as well as to foster the creation of structured, semantically-rich knowledge from paper contents [6]. Ubiquitous Knowledge Processing (UKP) Lab, at the Technische Universität Darmstadt, worked in two projects; first one is arguments extraction from scientific publications in the educational domain, where it should be possible to search the extracted arguments for a given topic and to retrieve related arguments for a given argument. The second one is Domain-Adaptive text mining to support knowledge discovery in the scientific historical literature using Dr. Inventor Multi–layer Scientific Corpus with NLP methods. Some semi-supervised machine learning techniques also used to explore the regular use of language in the scientific literature to understand it from metaphors found in those texts.

According to T. Nasukawa & T. Nagano [7], huge amount of accessible textual data that has been increasing rapidly may potentially contain a great wealth of knowledge. Consequently, huge amount of work is required in reading all the text and organizing their content as well. A text mining technology called Text Analysis and Knowledge Mining, shortly called TAKMI has been developed to overcome this counterintuitive situation, where focusing on the specific information in each document has been given less emphasize analyzing rather than what a large set of documents indicates as a whole. Jimmy Lin & Boris Katz [8] developed an open-domain question answering system ARANEA, that embraces two different views of the World Wide Web: as a heterogeneous collection of unorganized documents and also as well as a source of carefully analyzed and organized information about specific topics and the system takes advantage of Web data to answer repeated questions. The system, to take advantage of these different facets of the Web, integrates two different patterns of question answering: (i) knowledge annotation using semi structured database methods and (ii) knowledge mining based on redundancy techniques of statistics.

3. METHODOLOGY

In the building of the proposed hybrid knowledge-mining ecosystem the research team was able to implement certain aspects and analytics of IBM Watson's capabilities, including discovery, natural language classifier, natural language understanding and retrieve and rank. Successfully obtaining the desired result of building a commodity cluster. Establishing a local network with a dedicated Cisco Catalyst 500 switch is the first required step, then configure the bios of 15 Dell Optiplex 990 CPU's to PXE boot accordingly; this so they are capable to boot under Metal as a Service (MAAS) direction. These 15 Dell Optiplex 990 CPU's work with Ubuntu Operating system with Juju and OpenStack Autopilot [4] environments. MAAS server that is powered by Ubuntu was implemented, providing the research team and its users with the availability of accessing every single machines (nodes) of the cluster, check, alter, or update any desired detail or functionality. Juju Controller [4] which is open source and cost efficient application modeling tool that deploys, configures, scales and operates the software on clouds, was installed and perfectly implemented giving

support to the environment. OpenStack Autopilot is being installed and it will be ready to start executing some applications and visualizing power consumption, cluster performance, and data.

The Data has been collected by using an internally developed node.js platform that works as an API that gathers data from Google Scholar with an specific set of filters including topic and keywords; implementation of this code is crucial for understanding the basic concept of system. Data is then stored inside RethinkDB Database,which is then set for being moved to IBM BlueMix Cloud environment where it is properly analyzed with some application programming interface within IBM's Watson capabilities including: discovery, natural language classifier, natural language understanding and retrieve and rank.

3.1 Discovery:

This process adds a subjective search and content analytics engine to applications, which can classify patterns, direction, and actionable insights that help to make better decision. Structured and unstructured data with pre-enriched content is brought together securely in this service and it uses query language, which is simple, to reduce the need for manual filtering of results. By converting, normalizing, enriching unstructured data, it excerpts values from there. It is possible with Discovery service to build subjective, cloud-based exploration applications promptly that can unlock prosecutable insights that are hidden in unstructured data [9]. We used IBM Watson's API discovery to find needed data inside a Google scholar paper that were analyzed. It was providing an instruction using the API by conducting a keyword search inside a paragraph within a scholar paper.

3.2 Natural language classifier:

Performing the natural language classifier service to the cluster, which, returns a sentence or phrase in the best matching classes. For instance, while a researcher submits a question, the service gives back keys to the best matching answers or next actions for the application. The researcher, by providing a set of representative strings and a set of one or more correct classes for each training, actually creates a classifier instance. The new classifier can receive new queries or phrases and give back the top matches with a probability value for each match, after training [10]. The natural language classifier is implemented right after discovery, the API goes through the text or words to classify the data, and therefore scholar paper being analyzed are structurally classified. It is used because it understands the intent behind text and returns a corresponding classification.

3.3 Natural language understanding:

It is a subtopic of NLP. In unstructured text, the process tends to analyze text to excerpt metadata from content such as sentiment, concepts, relations, semantic roles entities, keywords, categories, and emotion [11]. The process tends to analyze text to excerpt metadata from it. Using NLU to analyze the large amounts of articles that have been retrieved. We derive the specific content and concepts that we wish to store in our cluster' database. This is the process by which our articles are decomposed and the valuable information like research question and results are extracted and stored to create the databases for our questions and answers.

3.4 Retrieve and rank (R&R):

By using machine learning algorithms and combined searches to detect "signals" in the data, this service R&R of IBM Watson helps users to find the most admissible information for their query. Here the word signal means keywords, most important relevant words or information. For instance, by using R&R, an experienced researcher can quickly find solutions from dense amount of data [12]. This is important for our project because being able to retrieve a specific set of information from a huge data set is crucial. R&R is the analysis process we use in order to retrieve the information that makes up our question and answer databases.

Using our scholar.js script, our data sets were obtaining using keyword "NASA", for this instance we stored close to 4.39 million articles in our cluster database. From there, our articles were then uploaded to IBM Bluemix where the analytical process began. Obtaining an IBM Bluemix account requires payment, unless sign up is done as a student, in which case you get 30 days free and can use all of their free services for the limited amount of time. Since we are a school of low resources, we chose to take the most affordable route.. This is why we created the cluster using commodity hardware and we chose to use as many free services as we could.

All these APIs met the criteria of being cost efficient, portable, and the system is able to utilize it by discovering actionable information from large sets of data, uses mathematical analysis to derive patterns and trends that exist in data. To deliver on system's security guarantee, our infrastructure runs on IBM secure cloud area and a modified version of the Apache Spark runtime system would be installed on master and worker nodes, to performs analysis and instrumentation of analytics programs before executing them. Developers load their data into the service, train a machine learning model based on known relevant results, then leverage this model to provide improved results to their end users based on their question or query. Table 1 shows the detailed configuration information about Hades and all other nodes.

Table 1: Node Configuration.

Node Name	Features/Purpose	RAM	Storage
Hades	MAAS/Master	8 GB	500 GB
Nyx1	Juju Controller	8 GB	500 GB
Jupiter2	OpenStack Autopilot	16 GB	500 GB
Client Nodes (12)	RethinkDB	84 GB	7.5 TB

4. RESULTS

The successful creation of the commodity cluster, Hades, began with the installation of MAAS on the master node. Successful Installation of Juju on a separate node created a Juju controller, which was essential to easily scale out the cluster horizontally. A third node was used to install OpenStack Autopilot, successfully automating every step of the deployment, management, operations and update process for the cloud. Once the Hades infrastructure was complete, the rest of the nodes were booted up with Ubuntu 16.04.2 LTS. The configuration in each of their BIOS was set to the correct time and booting sequence, successfully commissioning 11 nodes to our cluster. Totaling to 13 Disks, 84GB of RAM, and allowing for 7.5 TB of storage capacity. Figure 1 illustrates the Hades cluster created.

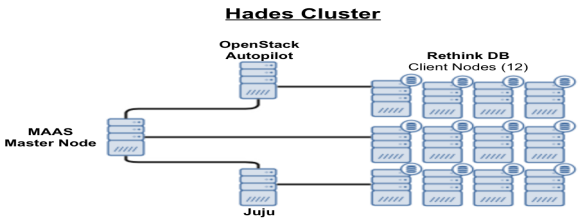

Figure 1. Hades Cluster

The use of node.js was essential to quickly deploy a fast and scalable network applications on server framework. node.js allowed us to successfully develop the code scholar.js internally, which pulls all indexed publications from Google-Scholar search. Data like: title, author, citation, url, summary and pdf are stored in the high-performance database called RethinkDB on cluster Hades. Using RethinkDB a Topics Database is created; from here data about the indexed publications are stored. Once configured, the local database will be able to efficiently extract, mine and analyze using distributed database as a service (DbaaS) named Cloudant NoSQL DB on IBM Bluemix.

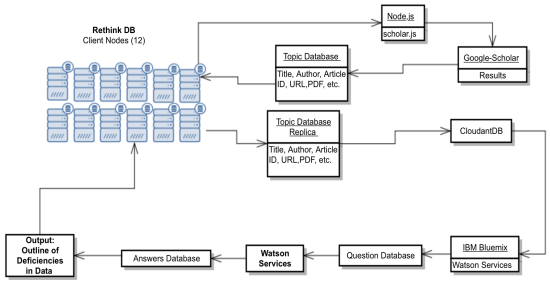

Figure 2. Analysis of Data

Figure 2 Illustrates the process through which our Topics Database, Questions Database, and Answers Database are used for analysis of our data.

The logical analysis of the big data is done on the IBM Bluemix cloud platform using the Watson Analytics services: discovery, natural language understanding, natural language classifier, and retrieve and rank. Using these services the data is analyzed in three separate instances. The first logical analysis is performed on the data stored in the Topics database, revealing research questions being asked in each of the indexed articles. This data is stored in the Questions database. The second analysis provides the final results for each of these research questions in the Questions Database; this data is stored in the Answers database. A third logical analysis, performed on the Answers database, results in an outline of the deficiencies in research data in that field. Yielding a new set of questions and answers expanding what is known of any one particular field.

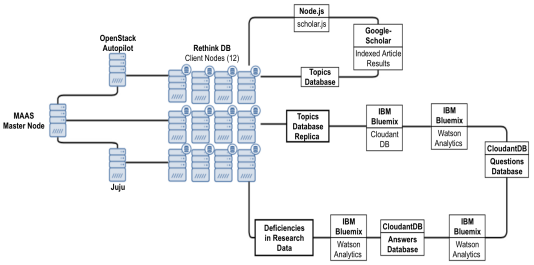

Figure 3. Process Map

Figure 3 depicts a simplified overview of our process from retrieval of scholarly results, to the logical analysis of our data, to the product, which outlines deficiencies in that data.

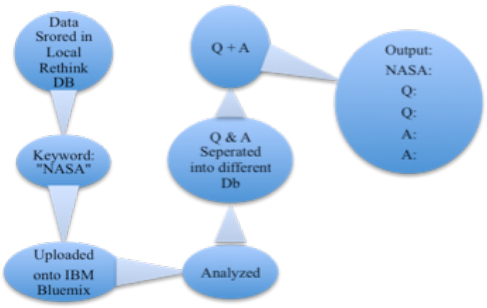

Figure 4. Expected Final Output

Figure 4 demonstrates the process through which our final output is expected.

The final output process can be broken down into two major steps. The first step is the process through which all of the articles, pertaining to the desired keyword. In this case, the keyword "NASA" returns approximately 4.39 million article results, these articles are then parsed and stored in our local cluster Rethink database for availability. The second step includes taking the readily available data and uploading onto the Cloudant database on the IBM Bluemix platform. An analysis is then performed using Watson's services natural language understanding and retrieve and rank. This breaks down each of the articles into two entities, a research question being asked and the answer that was found in that article. The research question and answers are then stored in their corresponding databases. A second analysis is done, using natural language classifier and retrieve & rank in order to take the analyzed questions and answers and convert them into a new document that outlines research areas and their deficiencies.

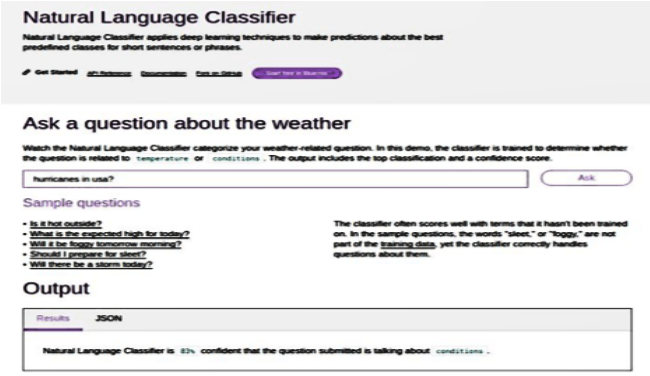

Figure 5. Natural language Classifier

Figure 5 explains how a user query matches our database by using natural language classifier where we used the weather as an example.

Figure 6. Sentiment Analysis

Figure 6 shows how Watson natural language understanding analyses the data with respect to sentiment analysis.

Also, emotion, keywords search and semantic rules analysis also could be done using natural language understanding that is depicted in figure 7,8 and 9 respectively. We used a sample NASA paper to extract the meta data from content.

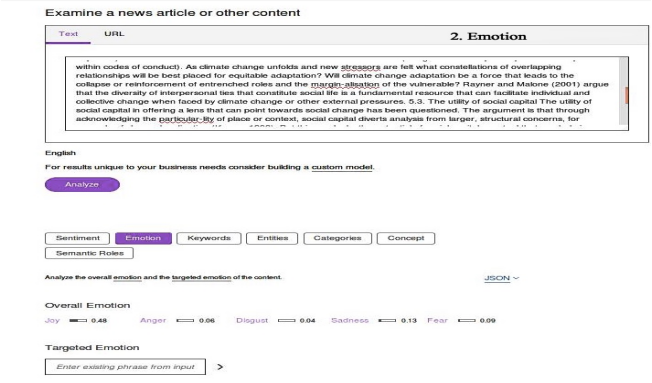

Figure 7. Emotion Analysis

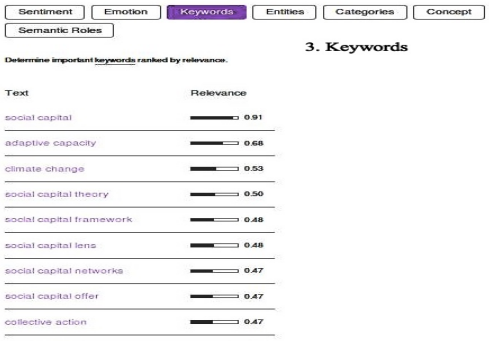

Figure 8. Keywords search

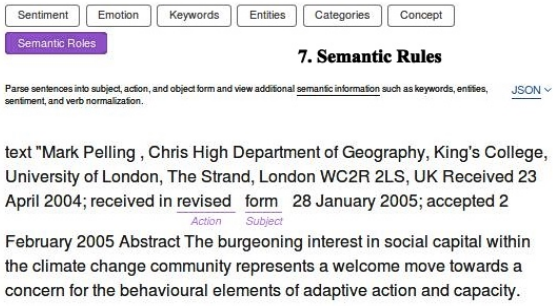

Figure 9. Semantic Rules

5. CONCLUSION

With the huge amount of scientific publications and documents publishing annually, we believe our cost-efficient Data Driven Secure Cloud Ecosystem will provide researchers with services that will help greatly to examine and operate data. We have described the design and implementation of a data driven secure cloud eco system in order to process, analyze and enrich the contents of a scientific article stored in our database. As future work, we plan to further improve our infrastructure to create a more detailed overview of the deficiencies in the research data for that field.

6. ACKNOWLEDGMENTS

Our special thanks goes to the Computer Science Department at Jackson State University for allowing us to use the campus resources and operating systems lab.

7. REFERENCES

[1] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, *et al.*, "Big data: The next frontier for innovation, competition, and productivity," 2011.
[2] E. Orduña-Malea, J. M. Ayllón, A. Martín-Martín, and E. D. López-Cózar, "About the size of Google Scholar: playing the numbers," *arXiv preprint* arXiv:1407.6239, 2014.
[3] K. M. Albert, "Open access: implications for scholarly publishing and medical libraries," Journal of the Medical Library Association, vol. 94, p. 253, 2006.
[4] S. Vemula, "Performance Evaluation of OpenStack Deployment Tools," ed, 2016.
[5] D. P. O'Donoghue, Y. Abgaz, D. Hurley, and F. Ronzano, "Stimulating and simulating creativity with Dr inventor," 2015.
[6] F. Ronzano and H. Saggion, "Knowledge extraction and modeling from scientific publications," in *International Workshop on Semantic, Analytics, Visualization*, 2016, pp. 11-25.
[7] T. Nasukawa and T. Nagano, "Text analysis and knowledge mining system," *IBM systems journal*, vol. 40, pp. 967-984, 2001.
[8] J. Lin and B. Katz, "Question answering from the web using knowledge annotation and knowledge mining techniques," in *Proceedings of the twelfth international conference on Information and knowledge management*, 2003, pp. 116-123.
[9] ibm.com. (2017). *Watson Documentation.* Available: https://www.ibm.com/watson/developercloud/discovery/api/v1/
[10] ibm.com. (2017). *Watson Documentation.* Available: https://www.ibm.com/watson/services/natural-language-classifier/
[11] ibm.com. (2017). *Watson Documentation.* Available: https://www.ibm.com/watson/services/natural-language-understanding/
[12] ibm.com. (2017). *Watson Documentation.* Available: https:///www.ibm.com/watson/services/retrieve-and-rank/