Integration and Analysis of Unstructured Data for Decision Making: Text Analytics Approach

Orobor Anderson Ise

Abstract— **Relational Database Management System (RDBMS) which is highly relied on by organizations for decision making are limited in their design to integrate and analyze data from unstructured sources. Research has shown that large part of organizational information exists in unstructured sources which might contain information needed for decision making. Integrating data from unstructured sources into RDBMS for the purpose of analysis is challenging due to their inconsistent and unorganized structures. This paper is therefore, aimed at developing a system that automatically integrates unstructured data into RDBMS. Considering the invaluable role played by academic journals (which are in turn unstructured in nature) in educational domain, the system, using text analytic approach, extract relevant information from academic journals to build a structured database which can further be analyzed to support decision making.**

Keywords— **Unstructured Data, Decision Making, Text Analytics, Conditional Random Field, Name Entity Recognition and Classification.**

I. Introduction

Business organizations strongly rely on a relational database for decision making or for business analysis [1]. Educational institutions in this regards are also being exhorted to support their decision based on data [2]. They had continually seek more efficient technology to better manage and support decision making procedures or assist them to set new strategies and plans [3]. Universities in Nigeria today generate massive data concerning its diverse fields of learning and other activities [4]. As the amount of data increases so do the challenges faced in trying to extract information from the ever growing data. Most of these universities had become data rich but information poor. This is because data at their disposal comes from different sources and are stored in different formats, locations, and schemas which pose a challenge in integration and analysis. The ability to carefully analyze these vast amounts of data makes it possible for universities to find pattern as well as gaining a better insight into their operations.

The challenges of integrating data become more complicated if the data is unstructured. Unlike structured data which are inherently record oriented and typically stored with a predefined schema that makes it very easy to query, search, analyze and integrate [5-7]. Unstructured data due to their nature is difficult to query, extract and integrate with other data sources. They do not easily fit into data-processing format or store in RDBMS, it remains stored, but unanalyzed [8]. Relying solely on structured data (which constitute small percentage of the organisation's data) is equivalent to making decision based on limited and incomplete information [9]. In the light of these challenges, there is a need to have a system capable of integrating and analyzing unstructured data since bulk (80% - 85%) of organizational data critical for decision making are unstructured [5-11].

This paper therefore, intend to address the above issue by demonstrating how unstructured data present in educational domain in form of academic journals can be integrated into RDBMS for further analysis to unlock useful information that could help improve decision making especially during promotion exercise for academic staff.

II. LITERATURES

A. Structured and Unstructured Data

Structured data generally refers to data that has a defined format [12] which makes it easy to query, analyze, and integrate with other structured data sources [1]. While unstructured data are those that do not conform to a specific, pre-defined data model [13]. Unstructured data consists of freeform text such as word processing documents, e-mail, Web pages, audio and video streams, images, and text files, as well as sources that contain natural language text [5]. The major obstacle in extracting information from such unstructured document is its unorganized, ambiguous and collapsed content. This requires a higher level of pre-processing rather than the typical preprocessing techniques adopted for structured and semi structured documents [7].

Enterprises are increasingly interested in accessing unstructured data and integrating it with structured data [5]. Unstructured data presents significant challenges for data scientists, often requiring an inordinate amount of time to structure and prepare data for analysis [14, 15]. Unstructured data sources are very high in volume. So, the challenges facing data are: getting the right information from it, transforming it into knowledge, analyzing it to find patterns and trends, storing information for fast and efficient access, managing the workflow and finally, making useful business intelligent reports [16].

Integrating data stored in both structured and unstructured formats can add significant value to an organization. Analyzing and processing unstructured data, formatting it and merging it with traditional structured data provides greater corporate insight for decision maker [8]. Knowledge may be discovered from many sources of

Manuscript received August 10, 2016. I wish to acknowledge the support of Prof. V.V.N Akwukwuma of Department of Computer Science, University of Benin. Integration and Analysis of Unstructured Data for Decision Making: Text Analytics Approach.

A.I. Orobor is with Federal University of Petroleum Resources Effurun, Delta State, Nigeria (corresponding author phone: +2347039190064; email: orobor.ise@gmail.com, orobor.anderson@fupre.edu.ng)

information, yet, unstructured texts remain the largest readily available source of knowledge [17].

B. Text Analytics

Text analytics is the process of analyzing unstructured text, extracting relevant information, and transforming it into structured information that can be leveraged in various ways [18]. According to [1], text analytics technique utilizes information retrieval and extraction and also uses natural language processing (NLP) for processing textual sources into models that support structured sources. Some of the tools used in text analytics includes: STANFORD CoreNLP, natural language toolkit (NLTK), general architecture for text engineering (GATE), APACHE OpenNLP and LingPipe. One of the oldest and best understood text analytics technologies is named entity extraction and is one of the core features of processing unstructured text [18, 19]. Successful information extraction depends on accurate named entity recognition which is the fundamental techniques of text analytics [20].

C. Name Entity Recognition and Classification (NERC)

Named entity recognition and classification (NERC) refer to the computational method to automatically recognize named entities in natural language text [22]. NERC involves identification and classification of proper names in texts into a set of pre-defined categories of interest. NERC identify and classify expression or every word/term in a document into some predefined categories [22, 23]. These expressions range from proper names of persons, organizations, locations or dates and often hold the key information in texts. NERC task was firstly introduced at MUC-6 in 1995; performance is around 90% for English and 70% for Czech [24]. The time precision, recall and F-measure (also F-score or F1 score) are used as a standard in NER evaluation metrics [24, 25]. These measures can generally be classified into two classes which are positive and negative having four classes of classification results:

1. Positive (P) or True Positive (TP) - positive object marked as positive.
2. False Positive (FP) - negative object marked as positive.
3. False Negative (FN) - positive object marked as negative.
4. Negative (N) - negative object marked as negative.

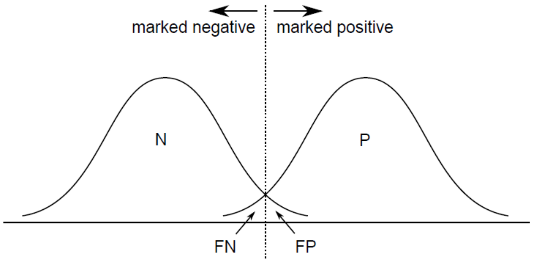


Figure 1. Precision and recall curve. Source [26].

Then the precision, recall and F-measure can be defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \qquad (1)$$

$$\text{Recall} = \frac{TP}{TP + FN} \qquad (2)$$

$$\text{F} - \text{measure} = \frac{2PR}{P + R} \qquad (3)$$

D. Conditional Random Field (CRF)

High performance NERC systems had often used conditional random fields (CRFs) as supervised machine learning method due to its relaxation on feature independence assumptions hence the advantage of handling high dimensional arbitrary feature sets over other machine learning methods such as hidden markov models (HMMs), maximum entropy markov models (MEMMs), and support vector machines (SVMs) [27, 28]. According to [28], a linear chain CRF defines a conditional probability. Let $x_{1:N}$ be the observations (e.g., words in a document), and $z_{1:N}$ the hidden labels (e.g., tags).

$$p(z_{1:N}|x_{1:N}) = \frac{1}{Z}\, exp\left(\sum_{n=1}^{N}\sum_{i=1}^{F} \lambda_i f_i(z_{n-1}, z_n, x_{1:N}, n)\right) \qquad (4)$$

Where, Z = partition function
F = weighted features
N = word positions
λ = parameters

In the model, scalar Z is the normalization factor, or *partition function*, to make it a valid probability and it defines the sum of exponential number of sequences.

$$z = \sum_{z1:N} exp\left(\sum_{n=1}^{N}\sum_{i=1}^{F} \lambda_i f_i(z_{n-1}, z_n, x_{1:N}, n)\right) \qquad (5)$$

The feature functions are the key components of CRF. The general form linear-chain CRF of a feature function is $f_i(z_{n-1}, z_n, x_{1:N}, n)$,which looks at a pair of adjacent states z_{n-1}, z_n, the whole input sequence $x_{1:N}$, and where we are in the sequence (n).

$$f_{1}(z_{n-1}, z_n, x_{1:N}, n) = \begin{cases} 1 & if\ z_n = PERSON\ and\ x_n = John \\ 0 & Otherwise \end{cases} \qquad (6)$$

The usage of feature depends on its corresponding weight 1. If 1 > 0, whenever f_1 is active (i.e. we see the word John in the sentence and we assign it tag PERSON), it increases the probability of the tag sequence $z_{1:N}$.

E. Related Works

[5] proposed an architecture that uses natural language processing and machine learning based techniques as a preprocessing step toward integrating structured and unstructured data. [8] work was on the use of text tagging and annotation, to derive business intelligence from business invoices of a company. The idea is extract information from

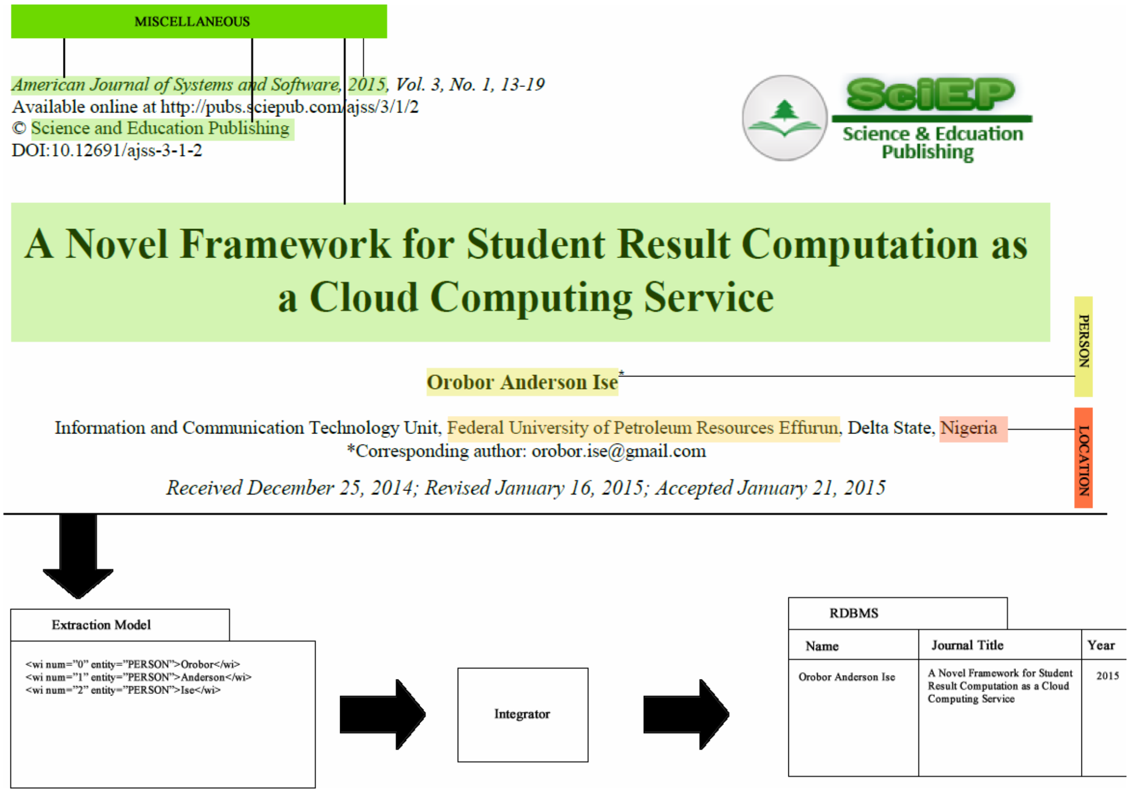
unstructured data in claim forms and linking it with an external knowledge base. [1] carried out a study on text analytics and its ability to transform textual sources to support structured environments to come up with a framework to deal with unstructured sources for decision making. Discovering knowledge from structured and unstructured data in customer relationship management (CRM) for effective decision making was the main focus of [6]. In their work, structured component is selected based on the resulting keywords from the unstructured text preprocessing process, and association rules is generated based on the modified generating association rules based on weighting scheme (GARW) algorithm. [29] presented how to organise and analyse textual data for extracting insightful customer intelligence from a large collection of documents and for using such information to improve business operations and performance. [11] describe an approach and system for managing and joining enterprise semi-structured data in a high-throughput, nimble, and scalable systems with traditional RDBMS. The paper presents the second release of NASA's NETMARK system. NETMARK is an enterprise information integration (EII) framework based on a modern schema-less concept approach. NETMARK schema-less information integration reinvents the way of managing semi-structured documents within traditional RDBMS. [30] presented a proposal to identify and extract concepts and named entities in legal documents. The proposed methodology uses a SVM classifier to associate concepts to legal documents and a natural language parser to identify named entities, namely, locations, organizations, dates, and references to other articles and documents. [31] in his work examined the incorporation of unstructured data from electronic clinical records for the task of classifying dementia progression status of subjects in a study on Alzheimer's disease, and additionally explored integration of these data and models with those of structured data.

III. MATERIALS AND METHODS

Unstructured data cannot simply be integrated and analyzed using RDBMS. Enterprise applications are now trying to bridge the gap of unstructured and structured data. This task requires extracting structured columns values from unstructured sources and mapping them to database entities. In this paper, we design unstructured data integration and analysis system that uses text analytic techniques in analyzing academic journals corpus in order to extract relevant information.

The fundamental architecture of the system is based on Stanford NER model. The system provides user with the capability of dragging and dropping journals into a specific area in the webpage where it is automatically preprocessed and integrated to the database. The task begins by transforming unstructured nature of journal into structured form that can easily fit into RDBMS. The process of transformation is in two phases. In phase 1, which is the preprocessing phase, structure is added to unstructured data using NER classifier. The output of this phase is in eXtensible Markup Language (XML) format which in turn act as an input to Phase 2 where the data is integrated to the RDBMS. As shown in figure 2 below.

The extracted entities which are enclosed within XML tags are predefined by the user during system configuration and distributed across the database entities.



American Journal of Systems and Software, 2015, *Vol. 3, No. 1, 13-19*
Available online at http://pubs.sciepub.com/ajss/3/1/2
© Science and Education Publishing
DOI:10.12691/ajss-3-1-2

A Novel Framework for Student Result Computation as a Cloud Computing Service

Orobor Anderson Ise*

Information and Communication Technology Unit, Federal University of Petroleum Resources Effurun, Delta State, Nigeria
*Corresponding author: orobor.ise@gmail.com

Received December 25, 2014; Revised January 16, 2015; Accepted January 21, 2015

```
<wi num="0" entity="PERSON">Orobor</wi>
<wi num="1" entity="PERSON">Anderson</wi>
<wi num="2" entity="PERSON">Ise</wi>
```

Name	Journal Title	Year
Orobor Anderson Ise	A Novel Framework for Student Result Computation as a Cloud Computing Service	2015

Figure 2. Unstructured data preprocessing and integration phases

A. Model for Extraction

Stanford NER, also known as CRFClassifier was adopted in our model for information extraction. CRFClassifier is a NER that labels sequence of words in a text. It comes with a well-engineered feature extractor for NER and many options for defining feature extractors. The classifier is included with 4 class model trained on the CoNLL 2003 eng.train, a 7 class model trained on the MUC-6 and MUC-7 training data sets, and a 3 class model trained on both dataset and some additional data including ACE 2002 [32]. The models are: 3 class: Location, Person, Organization, 4 class: Location, Person, Organization, Miscellaneous and 7 class: Location, Person, Organization, Money, Percent, Date, Time.

In this paper, we are interested in extracting data with predefined entities from academic journals. Applying CRFClassifier to our extractor model, we are able to extract the following entities: authors name, organisation, journal name, title, publisher, date accepted, date published, volume, year, keyword and abstract from journals as shown in table 1 BibTex format.

Table 1. Entity name extraction in BibTex format

```
Article{□
  author = ( Orobor Anderson Ise ),□
  title = ( A Novel Framework for Student Result Computation as a Cloud Computing Service ),
  journal = ( American Journal of Systems and Software ),
  publisher = ( Science and Education Publishing ),
  volume = ( 3 ),
  year = ( 2015 ),
  keywords = ( student result computation, cloud computing, result service, Nigeria university ),
  abstract = ( null )□
}
```

The problem with the task of entity extraction is that academic journals do not always maintain a particular structure depending on the publisher. An efficient model for data extraction should be able to handles these irregularities and the aim of this paper is to extract these entities irrespective of how they appear in the journal and map it to their corresponding entities in the database for storage as shown in figure 2.

We annotated 4 classes of named entities in academic journal corpus:

I. PERSON/PER: this entity type includes authors name and those appearing in the references section of the journal. For example Orobor Anderson Ise.

II. ORGANISATION/ORG: this entity type includes author's organisation as well as any other organisation that appears on the journal. For example Federal University of Petroleum Resources Effurun.

III. LOCATION/LOC: this includes geographical location of the author. For example Nigeria.

IV. MISCELLANEOUS/MISC: these are entities that do not belong to PERSON, ORGANISATION or LOCATION class. This type includes Title of Journal, Journal Name, and Publisher.

IV. RESULTS AND DISCUSSION

In this paper, our intention is to develop a system capable of transforming unstructured data in form of academic journals to structured form. The system using text analytics techniques based on natural language processing and machine learning technique is capable of:

1. Extracting specific data from unstructured academic journals.
2. Transform the data into structured data.
3. Integrate the structured data into RDBMS for further analysis.

The system is developed as a web based application, built on .NET framework using ASP.NET model view controller (MVC), Code First Workflow using C# programming language with the aid of Microsoft Visual Studio 2012. All data extracted from unstructured journals are stored using Microsoft SQL server. At the core of the system is CRFClassifier which enables the system to recognize entities and extract features from journals. All experiments were run on a desktop machine with Intel Core i3 processors running at 2.67GHz, 4GB RAM and Windows 7 operating system.

We experimented with 50 randomly selected articles from Google scholar containing 149,013 tokens of which 145,783 were words and the rest punctuations. The dataset contains 109 authors names, 18 of which were distinct. A sequence labeling CRF model found in StanfordCRF classifier were used to extract and annotate named entities as depicted in table 2.

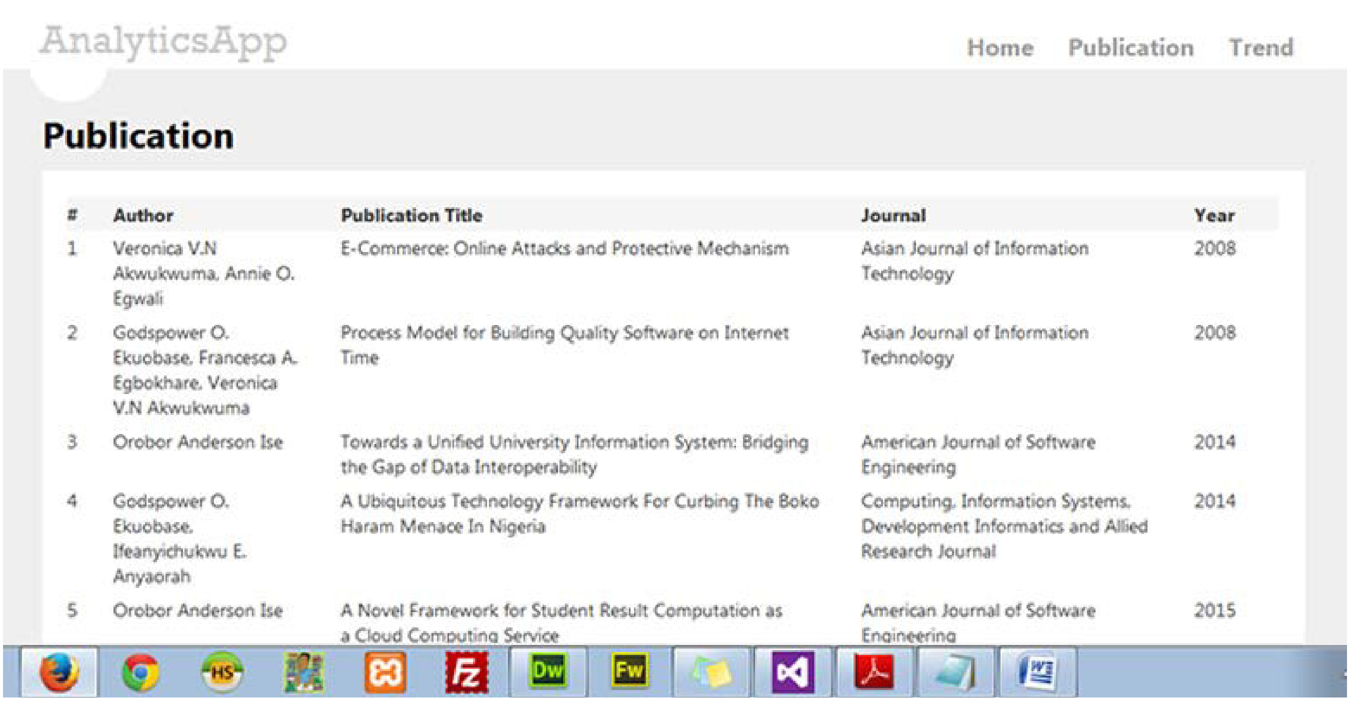
Table 2. Distribution of name entities across classes

Class	PER	ORG	LOC	MISC	TOTAL
Instances	109	43	9	651	812

On successful extraction of entities of interest in figure 1 using CRFClassifier, we could map the name entities to the database entities. The figure 3 below is a snapshot of the fully implemented data integration process.

The performance and accuracy of our extraction model is based on NER evaluation metrics: Precision (P), Recall (R) and F measure. Precision is the fraction of the retrieved documents that are relevant to the users information needs and is the ratio of correct matches to the total matches made, TP / (TP + FP). Recall is the fraction of the documents that are relevant to the query that is successfully retrieved and is the ratio of correct matches to the possible matches, TP / (TP + FN). F-Measure is the weighted harmonic mean of precision and recall. The result of the evaluation of each entities names integrated is presented in table 3.

The low precision and recall experienced in some of the classes is as a result of the complexity of some of the journals as well as missing name entities. Some journal includes names with abbreviations such as Orobor A.I and V.V.N Akwukwuma and also with wide diversity of organizations. Nearly every file that is positive is correctly identified as such with 79% recall in PERSON class. This means very few false negatives in the positive class. Positive classification with 74% for PERSON class is likely to be correct and so on.



Publication

#	Author	Publication Title	Journal	Year
1	Veronica V.N Akwukwuma, Annie O. Egwali	E-Commerce: Online Attacks and Protective Mechanism	Asian Journal of Information Technology	2008
2	Godspower O. Ekuobase, Francesca A. Egbokhare, Veronica V.N Akwukwuma	Process Model for Building Quality Software on Internet Time	Asian Journal of Information Technology	2008
3	Orobor Anderson Ise	Towards a Unified University Information System: Bridging the Gap of Data Interoperability	American Journal of Software Engineering	2014
4	Godspower O. Ekuobase, Ifeanyichukwu E. Anyaorah	A Ubiquitous Technology Framework For Curbing The Boko Haram Menace In Nigeria	Computing, Information Systems, Development Informatics and Allied Research Journal	2014
5	Orobor Anderson Ise	A Novel Framework for Student Result Computation as a Cloud Computing Service	American Journal of Software Engineering	2015

Figure 3. Screen shot of the extracted and integrated academic journals.

Table 3. Accuracy of name entities extraction

Class	Precision (%)	Recall (%)	F-Measure (%)
PER	74	85	79
ORG	71	74	72
LOC	74	68	71
MISC	85	74	79

This paper is of high significance to the educational institutions most especially the universities as:

1. It presented a simplified approach on to consolidation of unstructured data with structured once to provide how a unified view and description can be achieved.
2. This paper enables educational domain to discover hidden pattern in unstructured data by employing text analytic techniques.
3. It provides an approach that would enable data owners to treat data as data rather than segregating it as structured or unstructured.
4. It provides appropriate tools that could be used to improved decision making leveraging on unstructured data from academic journals.
5. The system developed is able to identify trends in journals collected over a period of time.

V. CONCLUSION

Integrating dataset across domains can provide data users with capability to find, access and analyze data based on their needs [33]. Traditional enterprise approaches such as creating data warehouse that organize and analyze data are clear solution for structure data that is well understood within the organizational boundary. However, when dealing with unstructured data, particularly large corpora of text, traditional database gives way. The challenge of getting the right information from it, transforming it into knowledge, analyzing it to find pattern and trend, stored for fast and efficient access and making useful business intelligent (BI) report are usually encountered [16]. Research has shown that information hidden or stored in unstructured data can play critical role in decision making. Organization stands to derive significant value to support operations if both structured and unstructured data can be integrated. This paper has demonstrated how unstructured data in the form of academic journals in education domain can be integrated to improve decision making. The system uses text analytic approach to bridge the gap between structured and unstructured data and capable of extracting set of entities of interest from unstructured academic journals and store them in a RDBMS for further analysis. Academic journals are a major medium through which research findings are published [34]. Considering the invaluable role played by academic journals in educational domain, our system extract relevant information from these journals to build knowledge based repository that can be used to support decision making using text analytic approach.

VI. FUTURE WORK

The unstructured data integration and analysis system developed in this paper requires user to manually drag and drop data to a specific area on the webpage for it to be processed. Rather than doing this, in our future work we will attempt to automate the data collection by leveraging on different journal portals such as Google Scholar and CiteSeer on the web. We also intend to improve the result of our precision and recall by automatically building corpus from academic journals supplied which will in turn serve as a dictionary support for the NER model.

REFERENCES

1. Prasad K. and Ramakrishna S. Text Analytics to Data Warehousing. *International Journal on Computer Science and Engineering*. **2010**, 2(6), 2201-2207.

2. Delavari, N., Phon-Amnuaisuk, S. and Beikzadeh, M. Data Mining Application in Higher Learning Institutions. *Informatics in Education International Journal*. **2008**, 7(1), 31-54.

3. Effective Decision Making in Higher Educational Institutions using Data Warehousing and Data Mining. Available online: http://www.ijcst.com/vol33/5/alok.pdf. (accessed on 12 June 2015).

4. Perceived Records Management Practice and Decision Making Among University Administrators in Nigeria. Library Philosophy and Practice. Available online: http://www.webpages.uidaho.edu/~mbolin/atulomah.htm. (accessed on 13 October 2015).

5. Integrating Structured and Unstructured Data Using Text Tagging and Annotation. Available online: http://www.bi-bestpractices.com/view-articles/4735. (accessed on 24 October 2015).

6. Fatudimu, I.T, Uwadia, C.O and Ayo, C.K. Improving Customer Relationship Management through Integrated Mining of Heterogeneous Data. *International Journal of Computer Theory and Engineering*. **2012**, 4(4), 518-522.

7. Exploration and Analysis of Unstructured Business Data using Text Analytics: A Study. Available online: http://www.ijetae.com/files/Volume5Issue7/IJETAE_0715_18.pdf. (accessed on 2 October 2015).

8. Gupta, V. and Rathore, N. Deriving Business Intelligence from Unstructured Data. *International Journal of Information and Computation Technology*. **2013**, 3(9), 971-976.

9. Sukumaran, S. Enterprise Infrastructure Scores over Islands of Applications for Information Management. Infosys SETLabs Briefings. **2005**, 3(4).

10. Management Update: Companies Should Align Their Structured and Unstructured Data. Available online: https://www.gartner.com/doc/470721?ref=ddisp. (accessed on 10 January 2016).

11. Managing Unstructured Data with Structured Legacy Systems. Available online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.154.5017&rep=rep1&type=pdf. (accessed on 24 October 2015).

12. Structured Data in a Big Data Environment, Dummies. Available online: http://www.dummies.com/how-to/content/structured-data-in-a-big-data-environment.html. (accessed on 10 January 2016).

13. Big Content: The Unstructured Side of Big Data. Available online: http://blogs.gartner.com/darin-stewart/2013/05/01/big-content-the-unstructured-side-of-big-data/. (accessed on 8 January 2016).

14. Skytree 15.2 Delivers Integrated Machine Learning to Unstructured Text Data. Available online: http://www.skytree.net/company/pr/skytree-15-2-delivers-integrated-machine-learning-to-unstructured-text-data/. (accessed on 24 December 2015).

15. Applications of Machine Learning through Unstructured Text Data. Available online: http://www.skytree.net/tag/unstructured-data/. (accessed on 24 December 2015).

16. Gupta, V. and Gosain, A. Tagging Facts and Dimensions in Unstructured Data. *International Conference on Electrical, Electronics and Computer Science Engineering (EECS)*. 1-6 May 2013

17. Gupta V. and Lehal, G. A Survey of Text Mining Techniques and Applications. *Journal of Emerging Technologies in Web Intelligence*. **2009**, 1(1), 60-65.

18. Text Analytics Beginner's Guide Extracting Meaning from Unstructured Data. Available online: http://www.angoss.com/wp-content/uploads/2013/04/eBook-Text-Analytics-Beginners-Guide.pdf. (accessed on 24 October 2015).

19. Entity Extraction – Lexalytics. Available online: https://www.lexalytics.com/content/whitepapers/Lexalytics-WP-Entity-Extraction.pdf. (accessed on 24 October 2015).

20. Wilcock, G. Introduction to Linguistic Annotation and Text Analytics. *Synthesis Lectures on Human Language Technologies*. **2009**, 2(1), 1-159

21. Disease Named Entity Recognition by Machine Learning Using Semantic Type of Metathesaurus. Available online: http://www.ijmlc.org/papers/367-C3012.pdf. (accessed: 10 December 2016).

22. Integrating unstructured data into relational databases. Available online: http://tangra.si.umich.edu/~radev/767w10/papers/Week06/TextRepresentation/Mansuri.pdf. (accessed on 30 October 2015).

23. Ekbal A. and Bandyopadhyay S. Bengali Named Entity Recognition using Support Vector Machine. In Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages. 51–58 October 2008.

24. Named Entity Recognition. PhD Study Report, University of West Bohemia. Available online: https://www.kiv.zcu.cz/site/documents/verejne/vyzkum/publikace/technicke-zpravy/2012/tr-2012-04.pdf. (accessed on 10 January 2016).

25. Nadeau, D. and Sekine, S. A survey of named entity recognition and classification. *Linguisticae Investigationes*. **2007**, 30(1), 3–26.

26. Manning, C., Mihai S., John, B., Finkel, J., Bethard, S. and McClosky, C. The Stanford CoreNLP Natural Language Processing Toolkit. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 55-60 May 2014

27. Conditional random fields: probabilistic models for segmenting and labeling sequence data. Available online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.26.803&rep=rep1&type=pdf. (accessed on 2 October 2015).

28. Conditional Random Fields. Available online: http://pages.cs.wisc.edu/~jerryzhu/cs769/CRF.pdf. (accessed on 2 October 2015).

29. Analysis of Unstructured Data: Applications of Text Analytics and Sentiment Mining. Available online: https://support.sas.com/resources/papers/proceedings14/1288-2014.pdf. (accessed on 2 October 2015).

30. Using linguistic information and machine learning techniques to identify entities from juridical documents.

Available online: http://www.morganclaypool.com/doi/abs/10.2200/S00196ED1V01Y200906AIM006?journalCode=aim. (accessed on 4 January 2016).

31. Mining and Integration of Structured and Unstructured Electronic Clinical Data for Dementia Detection. Thesis, Rochester Institute of Technology. Available online: http://scholarworks.rit.edu/cgi/viewcontent.cgi?article=9737&context=theses. (accessed on 24 October 2015).

32. Stanford Named Entity Recognizer. Available online: http://nlp.stanford.edu/software/CRF-NER.html. (accessed on 8 November 2015).

33. Hendler, J. Data Integration for Heterogenous Datasets. *Big Data*. **2014**, 2(4), 205–215.

34. The Development of Academic Journals in Institutions of Higher Learning in Kano State, Nigeria. Available online: http://www.webpages.uidaho.edu/~mbolin/ahmedmohammed.htm. (accessed on 29 January 2016).