Journal of Computer Sciences and Applications, 2017, Vol. 5, No. 2, 42-49
Available online at http://pubs.sciepub.com/jcsa/5/2/1

DOI:10.12691/jcsa-5-2-1

Characterisation of Academic Journal Publications Using Text Mining Techniques

Adebola K. Ojo*, Adesesan B. Adeyemo

Computer Science Department, University of Ibadan, Ibadan, Nigeria
*Corresponding author: adebola_ojo@yahoo.co.uk

Abstract The ever-growing volume of published academic journals and the implicit knowledge that can be derived from them has not fully enhanced knowledge development but rather resulted into information and cognitive overload. However, publication data are textual, unstructured and anomalous. Analysing such high dimensional data manually is time consuming and this has limited the ability to make projections and trends derivable from the patterns hidden in various publications. This study was designed to develop and use intelligent text mining techniques to characterise academic journal publications. Journals Scoring Criteria by nineteen rankers from 2001 to 2013 of 50th edition of Journal Quality List (JQL) were used as criteria for selecting the highly rated journals. The text-miner software developed was used to crawl and download the abstracts of papers and their bibliometric information from the articles selected from these journal articles. The datasets were transformed into structured data and cleaned using filtering and stemming algorithms. Thereafter, the data were grouped into series of word features based on bag of words document representation. The highly rated journals were clustered using Self-Organising Maps (SOM) method with attribute weights in each cluster.

Keywords: *highly rated journals, text mining, self-organising maps, filtering and stemming algorithms*

Cite This Article: Adebola K. Ojo, and Adesesan B. Adeyemo, "Characterisation of Academic Journal Publications Using Text Mining Techniques." *Journal of Computer Sciences and Applications*, vol. 5, no. 2 (2017): 42-49. doi: 10.12691/jcsa-5-2-1.

1. Introduction

Ranking of journals is widely used in academic circles in the evaluation of an academic journal's impact and quality. Journal rankings are intended to reflect the place of a journal within its field, and the prestige associated with it. Journal rankings can be used to evaluate the research impact of individual academics. Hence rather than measuring the impact of an academic's individual articles, universities and governments use the ranking of the journal as a proxy for the quality and impact of an academic's articles [1]. Some measures used in the ranking include impact factor, eigenfactor, scimago journal rank, h-index and expert survey. Recently, some journals were blacklisted in some institutions due to poor ratings. As a result of this, [2] produced a list called The Journal Quality List, which is a collation of journal rankings from a variety of sources. It is published primarily to assist academics to target papers at journals of an appropriate standard. The list was originally collated by the Bradford University School of Management (1997-2001). Since then, the list has been updated and extended periodically to keep it current. It contains rankings of different journals, and is used all over the world. There are more than 5000 downloads yearly by academics across the world, cited in different academic publications [3].

The Journal Quality List (JQL) comprises of academic journals in the following broad areas: Economics, Finance, Accounting, Management, and Marketing. The rankings for each journal include sources such as Foundation National pou l'Enseignment de la Geston des Entreprises (FNEGE), ESSEC Business School Paris 2013, Erasmus Research Institute of Management Journals Listing 2012, Cranfield University School of Management 2012, Agence d'evaluation de la recherché et de l'enseignement superieur (AERES) 2012, Verband der Hochshullehrer fur Betriebswirtschaft 2011, University of Queensland 2011, Danish Ministry ranking 2011, Centre National de la Recherche Scientifique 2011, Financial Times 45 Ranking 2010, British Association of Business Schools (ABS) Ranking 2010, Australian Business Deans Council 2010, Wirtschaftuniversitat Wien 2008, Aston University 2008, University of Queensland 2007, Hong Kong Baptist University School of Business 2005, British Journal of Management 2004, Verband der Hochschullenhrer fur Betriebswirtchaft 2003, and Wirtschaftuniversitat Wien 2001 [2].

For the purpose of this study, we are concentrating on highly rated journals in the JQL 2013 list and we shall be using text mining techniques to elicit hidden knowledge from these journals.

2. Literature Review

Data mining is the process of analysing data from different perspectives (large databases or Big Data) and summarizing it into useful and previously unknown

information for users [4,5]. It derives its name from the similarities between searching for valuable information in a large database and mining a mountain for a vein of valuable core, which is, transforming data dust to data 'gold' [6,7,8,9].

Text data mining is a natural extension of data mining [9], and follows steps similar to those in data mining. The qualitative difference in text mining, however, is that it processes data from natural language text rather than from structured databases of facts [10]. Companies use text mining software to draw out the occurrences and instances of key terms in large blocks of text, such as articles, Web pages, complaint forums, or Internet chat rooms and identify relationships among the attributes [11]. Often used as a preparatory step for data mining, text mining often translates unstructured text into a useable database-like format suitable for data mining for further and deeper analysis [12]. [7] also described text mining as an emerging technology that can be used to augment existing data in corporate databases by making unstructured text data available for analysis.

There exist some relationships between data mining, information retrieval, statistics, web mining, computational linguistics, natural language processing and text data mining. The problem of Knowledge Discovery from Text (KDT) [13] is to extract explicit and implicit concepts and semantic relations between concepts using Natural Language Processing (NLP) techniques. Its aim is to get insights into large quantities of text data. KDT, while deeply rooted in NLP, draws methods from Statistics, machine learning, reasoning, information extraction, knowledge management, and others for its discovery process. KDT plays an increasing role in emerging applications, such as Text Understanding [14]. Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers). The documents retrieved should be relevant to the information needs of the user who performed the search query.

If the set of documents relevant to a query is denoted as {Relevant}, and the set of documents retrieved is denoted as {Retrieved}, then the set of documents that are both relevant and retrieved is denoted as {Relevant} ∩ {Retrieved}. The two basic measures for assessing the quality of text retrieval [15] are Precision and Recall. Precision is the percentage of retrieved documents that are in fact relevant to the query. Recall is the percentage of documents that are relevant to the query and were, in fact, retrieved. It is formally defined as

Text Mining is a non-traditional information retrieval (IR) method whose goal is to reduce the effort required of users to obtain useful information from large computerized text data sources. Traditional Information Retrieval often simultaneously retrieves both "too little" information and "too much" text [16,17]. However, in Information Retrieval (otherwise known as Information Access), no genuinely new information is found. The desired information merely coexists with other valid pieces of information.

Natural Language Processing (NLP) is a sub field of Artificial Intelligence (AI) and linguistics regions [18]. In NLP, Text Mining applications are also quite frequent and they are characterized by multilingualism [19]. Use of Text Mining techniques to identify and analyze web pages published in different languages, is one of its examples [14]. The main aim of NLP studying is the generation and realizing of natural languages. One direction of NLP research relies on statistical techniques, typically involving the processing of words found in texts [20]. One of the NLP applications in text retrieval is usage of these techniques as a necessary component in web search engines, via automated translation tools or in summary generators [21].

NLP techniques are used for text that is typically syntactically parsed using information from a formal grammar and a lexicon, the resulting information is then interpreted semantically and used to extract information about what was said [22]. It includes techniques like word stemming (removing suffixes) or a related technique, lemmatization (replacing an inflicted word with its base form), multiword phrase grouping, synonym normalization, part-of-speech (POS) tagging (such as elaborations on noun, verb, preposition), word-sense disambiguation, anaphora resolution and role determination (such as subject and object) [18,21].

The difference between regular data mining [23] and text mining is that in text mining the patterns are extracted from natural language texts rather than from structured databases of facts. Text mining tries to apply these same techniques of Data mining to unstructured text databases. To do so, it relies heavily on technology from the sciences of Natural Language Processing (NLP), and Machine Learning to automatically collect statistics and infer structure and meaning in otherwise unstructured text. The usual approach involves identifying and extracting key features from the text that can be used as the data and dimensions for analysis. This process is called feature extraction, is a crucial step in text mining.

Web mining [24,25,26] is the activity of identifying patterns implied in large document collection. Web mining is an integrated technology in which several research fields are involved, such as data mining, computational linguistics, statistics, and informatics. There is no generally acceptable definition of Web Mining. Since web mining derives from data mining, its definition is similar to the well-known definition of data mining [19]. Nevertheless, Web mining has many unique characteristics compared with data mining.

Text-mining is ideally suited to extract concepts out of large amounts of text for a meaningful analysis. It has been used in a wide variety of settings, ranging from biomedical applications to marketing and emotional/sentiment research where a lot of data needs to be analysed in order to extract core concepts. Text-mining achieves this, by applying techniques from information retrieval (such as Google), natural language processing, including speech tagging and grammatical analysis, information extraction, such as term extraction and named-entity recognition and data mining techniques, such as pattern identification [27,28].

Applications of text mining methods are diverse and include Bioinformatics [29], Customer profile analysis, Anti-Spam Filtering of Emails, Event tracks, Text Classification for News Agencies [30] and Web Search [31]. These applications also extend to any sector where text documents exist. For instance, history and sociology

researchers can benefit from the discovery of repeated patterns and links between events, *crime detection* can profit by the identification of similarities between one crime and another [32], and unsuspected facts found in documents may be used in order to populate and update scientific database [33]. Other areas include updating automatically a *calendar* by extracting data from *e-mails* [33,34], identifying the original source of a *news article* [35] and monitoring inconsistencies between *databases and literature* [36]; biomedical applications (for example, identification of biological entities, automatic extraction of protein interactions and associations of proteins to functional concepts); marketing applications (customer relationship management), and sentiment analysis [37].

3. Methodology

In this section, the processes that were involved in this study were discussed. The raw texts in electronic format were extracted from the abstracts of academic journal publications. These electronic resources were downloaded from online digital data sources. These raw texts were pre-processed and transformed. The features were extracted from the transformed texts and converted to form structured data. These structured data were further analysed and the results interpreted for Knowledge Management purposes (decision making). Figure 1 presents the logical text mining operations carried out in this study.

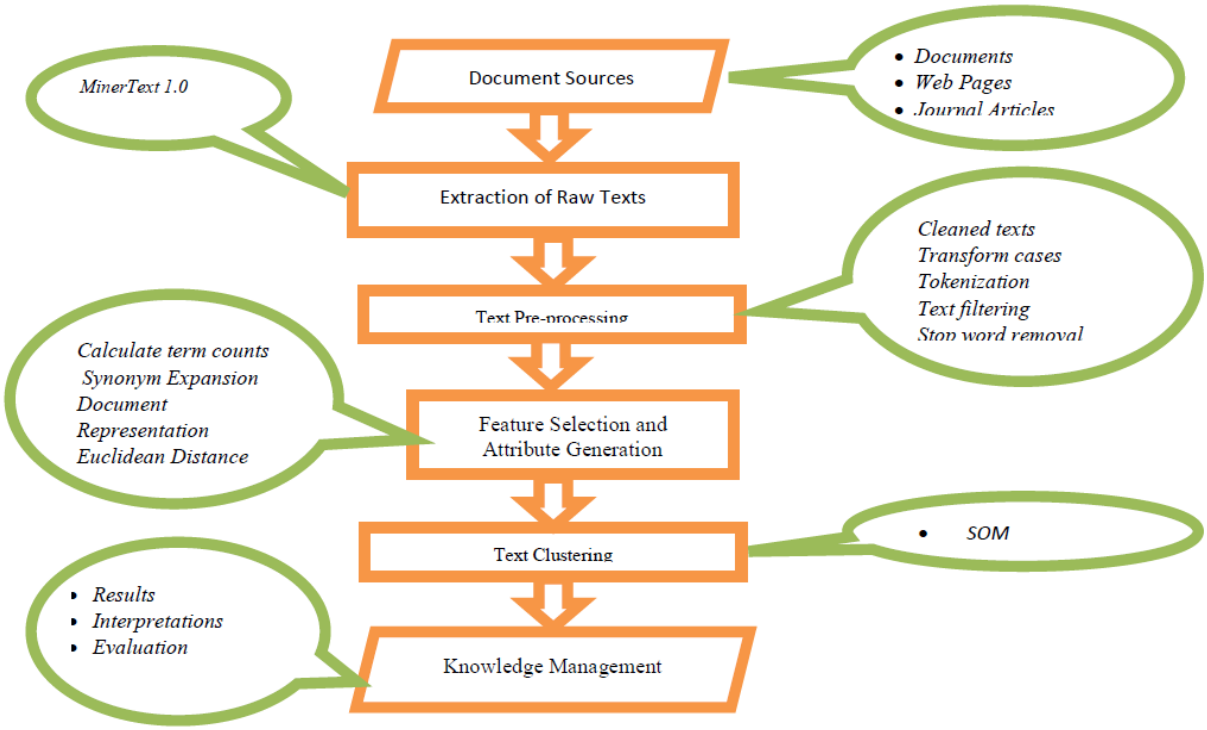

Figure 1. The Workflow of the System

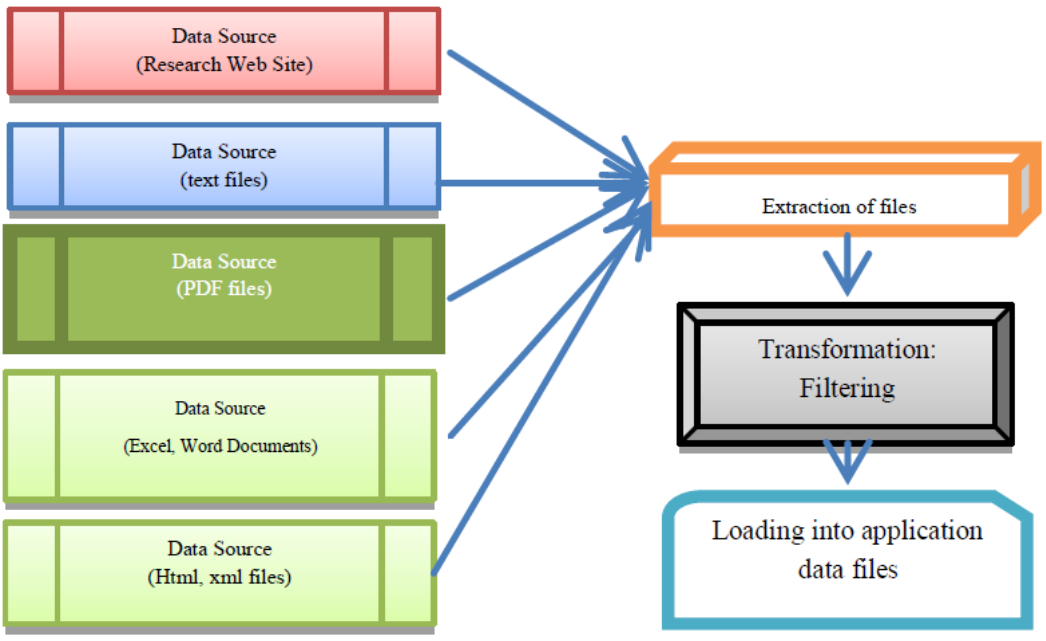

Figure 2. Different Data Sources

The data sources were electronic resources from the web. These were in form of text, CSV, doc, xls and html documents/files. These data were transformed and loaded into the application data files as shown in Figure 2.

3.1. Text Data Collection

The text data used were downloaded from journals in the Journal Quality List. The text data was imported into Microsoft Excel (CSV) format. Journal Quality List (JQL) is a collation of journal rankings from a variety of sources which is published primarily to assist academia to target papers at journals of an appropriate standard [2]. The list was originally collated while the editor was associated with the Bradford University School of Management (1997-2001). Since then, the list has been updated and extended periodically to keep it current. The 50th version of the JQL contained 23 different rankings of 940 journals. It comprised of academic journals in different areas such as Economics, Finance, Accounting, Management, Marketing, Business Administration/Management, Commerce, Education, Science and Technology, International Relations, Business Management, Medicine, Public Health, Arts and Humanities, History, Information Processing/Science, Knowledge Management, Engineering.

3.2. Text Data Selection

Article abstracts, authors' bibliometrics (authors' affiliations) and keywords of the case studies were used for the analysis. The text data for this research was obtained from journal publications ranked in the Journal Quality List (JQL). As at 2013, there were 50 editions of the JQL published [2]. The 50th edition contained the ranking of journals from 2001 to 2013. The journals cited in the edition were used in this study. Some journals in this edition were rated high. A non-probability method known as Purposive Sampling was used in selecting the journals from the JQL. Purposive sampling is a sampling technique whose elementary units are chosen according to the discretion of the expert who is familiar with the relevant characteristics of the population.

3.3. Text Pre Processing

This stage involved the use of information extraction tools to analyse the unstructured data by identifying all the key features within the text. Usually the abstracts, authors' bibliometrics and keywords provided enough semantic information that the journals used. These information were extracted from the full text and parsed into independent sentences. This was implemented using the text mining software (MinerText 1.0) that was developed for this study.

In text pre-processing, the document was first split into a series of words (features). Adjectives, adverbs, nouns and multi-word were extracted from the document. Word frequency and inverse document frequency were two parameters used in filtering terms. Low term frequency (TF) and document frequency (DF) terms were often removed from the indexing of journal documents. To better match concepts among terms, words were stemmed based on Porter's algorithm. It contained keywords, title words, and clue words.

In "Bags of words" representation each word was represented as a separate variable having numeric weight. The most popular weighting schema is normalized word frequency *tfidf*. The tf-idf (term frequency-inverse document frequency) statistic was based on the frequency of a given term in the record. This was normalized by being divided by the total number of times term appeared in all records.

$$f - idf = \frac{tf}{Df}. \quad (1)$$

Also,

$$fidf\,(w) = tf.\log\left(\frac{N}{df\,(w)}\right) \quad (2)$$

where
idf is the inverse document frequency
tf is the term frequency (number of word occurrences in a document)
df(w) represents document frequency (number of documents containing the word)
N gives the number of all documents
tfidf(w) is the relative importance of the word in the document

The text extraction involved the identification and extraction of texts from the scientific publications. Text characterisation involved the following transformation processes:

Transform Cases: In this stage, all characters (tokens) in the documents were transformed into lower cases respectively.

Tokenization: The language used in this study was Part of Speech (POS) tagger. Instead of using the complex methods, the tokenization process was accomplished by using space to split the sequence of characters into a sequence of tokens using non-letter character. This resulted in tokens consisting of one single word, which was used in building the word vector.

Filter Tokens (by Length): The tokens were filtered based on their length (that is, the number of characters they contained) which involved the minimum and maximum numbers of characters contained in each token. This was done by splitting the text of a document into a sequence of tokens. Minimum of two characters and maximum of twenty-five characters were used in filtering the total length.

Stop Word Removal: Extremely common words which would appear to be of little value in helping select documents matching a user's need were excluded from the vocabulary entirely. English stop words were filtered from the document by removing every token, which equalled a stop word from the built-in stop word list. These words were stop words and the process was called stop word removal. Such words included a, an, the, and the prepositions. Tokens contained in the stop word list were discarded.

Feature Selection and Attribute Generation

In this stage, a subset of the features was selected to represent a document. This created an improved text representation since many features had little information content. Stop Words were removed, and words stemmed down to their roots. Stemming identified a word by its

root and reduced dimensionality (number of features). Features were selected based on classification and some irrelevant attributes were removed.

Document Representation: Vector Space model is one of the efficient methods of representing documents as vectors using the term frequency weighting scheme. The entire data collection from the PDF/XML file was represented as vectors using the Vector Space Model.

Euclidean Distance: The cosine measure is a similarity measure rather than a distance. Distances are more comfortable to work with. Similarities were easily converted to distances using (3) by organising similarities into a positive-definite matrix *C*, where *ij*-th element of this matrix indicated the similarity of the *i*-th and *j*-th documents.

$$d_{ij} = \sqrt{c_{ii} - 2c_{ij} + c_{jj}}. \quad (3)$$

When two documents were the same ($c_{ii} = c_{jj}$) then the distance was zero.

Alternatively,

$$d_{ij} = \left(\sum_{k=1}^{m} \left(x_{i,k} - x_{j,k} \right)^2 \right)^{1/2} \quad (4)$$

i, j = records; m = number of variables; *x*= matrix of the *i*-th and *j*-th documents.

3.4. Text Clustering

The clustering algorithm used for this study was Self-Organising Maps (SOM). SOM is good in clustering because of the highest weights attributed to words in the clusters; besides it suggested more words that were used for classification in the journals.

The procedure of SOM for text clustering was summarized as follows:

i. Each node's weight was initialized.

ii. A vector was chosen at random from the set of training data and presented to the SOM network.

iii. Every node in the network was examined to calculate which ones' weights were most like the input vector. The winning node was commonly known as the Best Matching Unit (BMU), Using Euclidian Distance:

$$Dist\ From\ Input^2 = \sum_{i=0}^{i=n} \left(I_i - W_i \right)^2 \quad (5)$$

I = current input vector
W = node's weight vector
n = number of weights

iv. Adapting the vectors of the winner and its neighbors using (6):

$$m_i(t+1) = m_i(t) + \alpha(t).h_i(t).\left[x(t) - m_i(t) \right] \quad (6)$$

Where
mi (t+1) = neuron vector after adaptation
mi (t) = original vector of neuron I (neuron vector before adaptation)
α(t) = learning rate
hi (t) = neighbour rate
x(t) = document vector
t =time
[x(t) - mi (t)] = distance between neuron vector and document vector.

v. Step (ii) was repeated for N iterations.

4. Results and Discussions

Table 1 presents the summary of the seven highly rated journals selected by 19 rankers from 2001 to 2013, with their impact factors, total number of volumes and their total number of issues in each volume. The total number of issues of all the highly rated journals from inception till July 2013 was 1149 issues. Simple random sampling was used for this study in selecting the journal articles with the sample size of 10%. These were analysed as Academy of Management Review (AMR): 15; Accounting Review (AR): 7; Administrative Science Quarterly (ASQ): 6; Journal of Finance (JOF): 34; Journal of MIS Quarterly (MIS): 15; Journal of Marketing (JM): 5, and Strategic Management Journal (SMJ): 33. The total issues were 115.

The abstracts (text data) of the assessed journals were converted to word features. The stop words in the text were removed, and the resulting word data were stemmed down to their roots. Stemming identifies a word by its root and reduces dimensionality (number of features). Features were then selected based on classification. Some irrelevant attributes were removed. The abstracts (text data) were converted to word features.

The word dictionary constructed for highly rated journals is presented in Table 2.

The list of words was generated using bag of words representation and the number of times each word appeared in the documents, as shown in Figure 3.

Table 1. Highly Rated Journals

S/N	Journal	Year Started (Inception)	Total Volumes as at July 2013	Publications Per year	Total Issues as at July 2013	Impact Factor as at July 2013
1.	AMR	1976	38	4	152	7.895
2.	AR	1926	88	4, 6	72	1.92
3.	ASQ	1999	58	4	62	4.182
4.	JOF	1946	68	4, 6	341	4.333
5.	JM	1936	77	4, 6	46	5.47
6.	MIS	1977	37	Quarterly	150	4.659
7.	SMJ	1980	34	Varies (4…13)	326	3.367
Grand Total					**1,149**	

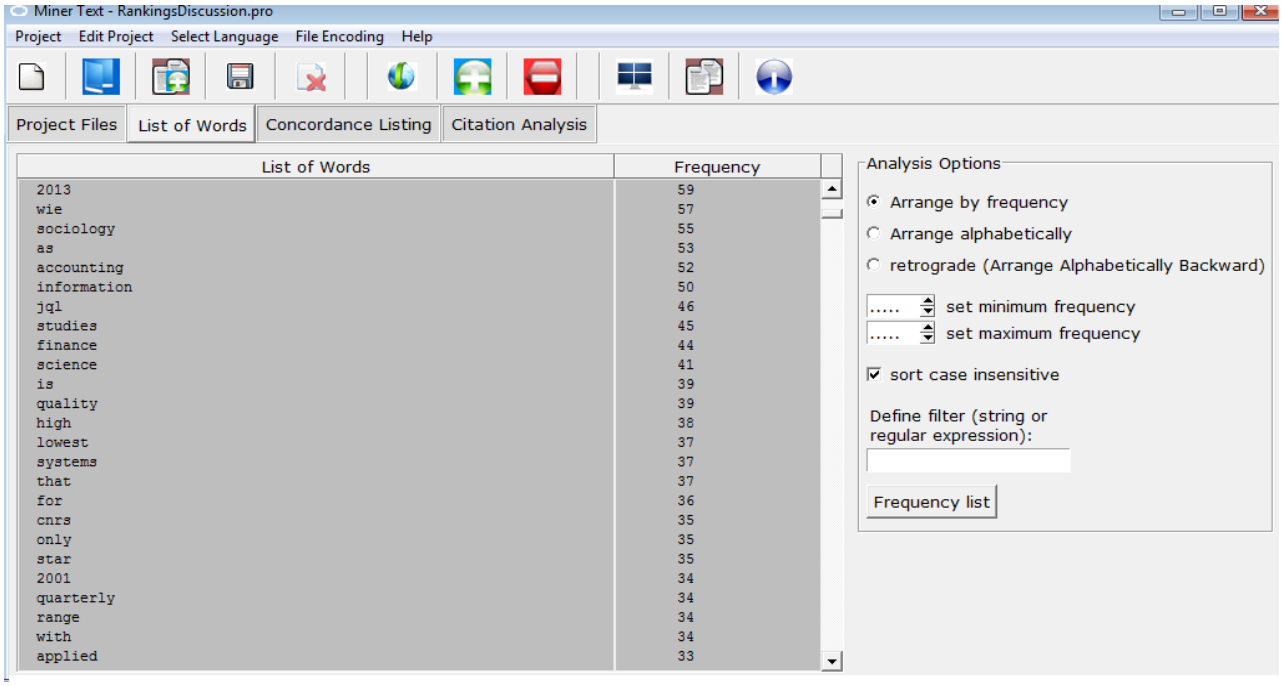

Figure 3. Text Pre-processing

Table 2. Data Dictionary for Highly Rated Journals

Data Attributes	Code	Code Description	Number of Word Features
Institution (I)	I1	University	768
	I2	Polytechnic	2
	I3	College	152
Designation (D)	D1	Professor	267
	D2	Assistant Professor	72
	D3	Lecturer I and II	49
	D4	Director/Dean	16
Faculty (F)	F1	Business Administration	224
	F2	Science and Technology	156
	F3	Economics	100
	F4	Finance	258
	F5	Marketing	244
	F6	Management	341
	F7	Commerce	2
	F8	Education	98
Location (L)	L1	North America	1825
	L2	South America	95
	L3	Asia	324
	L4	Australia	50
	L5	Europe	781
	L6	Africa	38

Table 3. Word Features in Highly Rated Journals

HRJ Journals	L	I	D	F
AMR	100	224	215	451
ASQ	103	46	0	9
AR	439	128	0	63
SMJ	666	135	77	365
MIS	528	3	6	220
JOF	1031	386	106	215
JM	246	0	0	100

Table 3 also shows the features that were generated.

Table 4 presents the clusters obtained by clustering the highly rated journal data by Institutions, Designations, Faculties and Locations.

In Table 4, I1, I2, I3 represent institutions which are University, Polytechnic and Colleges of schools, respectively. D1, D2, D3 and D4 represent Designations (or status of academic staff), that is, Professor, Assistant/Associate Professor, Lecturer I/II and Director. F1, F2, F3, F4, F5, F6, F7 and F8 represent Faculty, which are Business Administration, Science and Technology, Economics, Finance, Marketing, Management, Commerce and Education. Locations are represented by L1, L2, L3, L4, L5 and L6. L1 as North America, L2 as South America, L3 as Asia, L4 as Australia, L5 as Europe and L6 as Africa.

In cluster one, it was observed that most of the publications were from the universities and the highest number of publications came from North America. Cluster two shows that most publications were mostly Management journals, and majority of the publications were from North America. Cluster three revealed that the highest number of publications came from North America. Also, most of the publications were university publications. Cluster four showed that most publications came from North America and were university publications. Most of the dominant authors were Professors, and were mostly management journals. It was observed in cluster five that most of authors were Professors from universities. Most of these publications were from North America. They were mostly Business Administrative journals. In cluster six, most publication areas that dominated were from Finance. They originated from universities and were mostly from North America. Cluster seven revealed that, most of these publications were Management journals and were from the universities.

Table 4. Publication Clusters

	c 1	c 2	c 3	c 4	c 5	c 6	c 7
I1	0.4797	0.01	0.3774	0.3224	0.2119	0.0521	0.0111
I2	0	0	0	0	0	0	0
I3	0.0078	0.0028	0.0133	0.0969	0.0387	0	0
D1	0.0045	0	0	0.2462	0.2193	0	0
D2	0.0083	0	0.0282	0.2210	0.1312	0	0.0042
D3	0	0	0	0	0	0	0
D4	0	0	0	0.0357	0.0155	0	0
F1	0.0128	0.0120	0.0044	0.3110	0.2350	0.0103	0.0092
F2	0.1054	0.1132	0.0713	0.0117	0.0190	0.0138	0.0151
F3	0.0061	0	0	0.0092	0.0111	0	0.021874
F4	0.2316	0	0	0	0.0062	0.1949	0.0309
F5	0	0	0	0	0	0	0
F6	0.0323	0.3023	0.0933	0.2936	0.0115	0.0119	0.1849
F7	0	0	0	0.0059	0.0037	0	0.0034
F8	0	0	0	0	0	0	0
F9	0	0	0	0	0	0	0
L1	0.4456	0.1624	0.3507	0.3043	0.2306	0.0161	0.0114
L2	0	0.0071	0.0117	0	0	0	0
L3	0.0603	0.0102	0.0195	0.0054	0.02398	0	0.0088
L4	0.0075	0	0	0	0	0	0
L5	0.0330	0.0233	0.0196	0.0072	0.0109	0	0.0114
L6	0.0069	0	0	0	0	0	0

5. Conclusions

In this study, we focused on the characterization of academic journal articles using text mining techniques. This was done by using the developed text mining software to capture abstracts of academic publications from highly rated journals of Journal Quality List. These abstracts comprised of several issues with many articles. The raw article documents were split into a series of words (features). Stop words were removed, and words stemmed down to their roots, thereby transforming the unstructured data into structured data. The processed data 'mined' in this study identified patterns and extracted valuable information and new knowledge. The data of highly rated were classified into Institution, Location, Designation and Faculty. Self-Organizing Maps (SOM) clustering algorithm was used to also cluster the data. SOM was better in clustering because of the highest weights attributed to words in the clusters; besides it suggested more words that were used for classification in the journals. It was also able to depict texts in more figurative and better visual way. The model developed in this study is invaluable and useful for analysing and discovering patterns in academic electronic resources. It has also helped in identifying and characterizing the most relevant factors or features for determining how academic journals are ranked in academic institutions.

References

[1] Harzing, A.W. 2010 The Publish or Perish Book: Your guide to effective and responsible citation analysis, Melbourne: Tarma Software Research. http://www.harzing.com/publications/publish-or-perish-book

[2] Harzing, A.W. 2013. Journal Quality List. University of Melbourne Department of Management Parkville Campus Parkville VIC 3010 Australia. https://www.harzing.com/resources/journal-quality-list.

[3] Harzing, A.W. 2006. Journal Quality List. University of Melbourne Department of Management Parkville Campus Parkville VIC 3010 Australia. https://www.harzing.com/resources/journal-quality-list.

[4] Wu, Sheng-Tang. 2007. Knowledge Discovery Using Pattern Taxonomy Model in Text Mining. A dissertation in the Faculty of Information Technology, Queensland University of Technology. http://eprints.qut.edu.au/16675/1/Sheng-Tang_Wu_Thesis.pdf.

[5] Devedzic V. 2001. Knowledge discovery and data mining in databases. In S. K. Chang, editor, Handbook of Software Engineering and Knowledge Engineering, Vol. 1 – Fundamentals, pages 615-637. *World Scientific Publishing Co*, 2001.

[6] Osofisan, A. O. 2011. Transforming Data Dust to Data Gold. An Inaugural Lecture Delivered at the University of Ibadan, 25 August, 2011. Ibadan University Press.

[7] Francis, Louise and Flynn, Matt. 2010. Text Mining Handbook. Casualty Actuarial Society *E-Forum*

[8] Dorre J. Gerstl P. and Seiffert R. 1999. Text Mining: Finding Nuggets in Mountains of Textual data. *In Proc. 5th ACM Int. Conf. on Knowledge Discovery and Data Mining (KDD-99)*, pages 398-401, San Diego, US, 1999. ACM Press, New York, US.

[9] Hearst M., 1999. "Untangling Text Data Mining," *In the Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*.

[10] Chen, Kuan C. 2009. Text Mining e-Complaints Data From e-Auction Store With Implications For Internet Marketing Research Purdue University Calumet, USA. Journal of Business & Economics Research – May, 2009 Volume 7, Number 5

[11] Robb Drew. Taming Text. 2005. http://vnweb.hwwilsonweb.com/hww/jumpstart.jhtml?recid=0bc05f7a67b1790e8bd354a88a41ad89a928d23360302a4959035699f17e2ba8a63e2dd032c73f8a7fmt=H.

[12] Cerrito, Patricia 2005. Inside Text Mining. http://wilsontxt.hwwilson.com/pdffull/06619/275n6/gs9.pdf, March 24, 2005.

[13] Haralampos, Karanikas and Babis, Theodoulidis Manchester. 2001. Knowledge Discovery in Text and Text Mining Software, Centre for Research in Information Management, UK.

[14] Gupta, Vishal and Lehal, Gurpreet S. 2009. A Survey of Text Mining Techniques and Applications. *Journal of Emerging Technologies in Web Intelligence*, Vol. 1, No.1. August 2009.

[15] Umajancy. S. and Thanamani , Antony Selvadoss. 2013 An Analysis on Text Mining –Text Retrieval and Text Extraction, *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 2, Issue 8.

[16] Humphreys, K., Demetriou, G., and Gaizauskas, R. 2000. Bioinformatics applications of information extraction for scientific journal articles. *Journal of Information Science*. 26. 75-85.

[17] Sharp, K. 2001. Internet librarianship: Traditional roles in a new environment. IFLA Journal, 27 (2): 78-81.

[18] Gharehchopogh, F. and Z. Khalifelu, 2011. Analysis and evaluation of unstructured data: Text mining versus natural language processing. Proceedings of the 5th International Conference on Application of Information and Communication Technologies, Oct. 12-14, IEEE Xplore Press, pp: 1-4.

[19] Bolasco, S., Baiocchi, F., Canzonetti, A., della Ratta-Rinaldi, F., Feldman, A. 2004. Applications, sectors and strategies of Text Mining: a first overall picture, in S. Sirmakessi (ed.) Text mining and its Applications, *Springer* Verlag, Heidelberg, pp. 37-52.

[20] Manning, Christopher D. and Schiitze, Hinrich. 1999. Foundations of Statistical Natural Language Processing. The MIT Press Cambridge, Massachusetts London, England ics.upjs.sk/~pero/web/documents/pillar/Manning_Schuetze_StatisticalNLP.pdf.

[21] Baeza-Yates, Ricardo and Ribeiro-Neto, Berthier. 1999. Modern Information Retrieval. *ACM Press*, New York. Addison-Wesley. Retrieved from people.ischool.berkeley.edu/~hearst/irbook/print/chap10.pdf.

[22] Kao, A., and Poteet, S., 2004. Report on KDD Conference 2004 Panel Discussion Can Natural Language Processing Help Text Mining? SIGKDD Exp. Newsl. 6(2), 132-133.

[23] Navathe Shamkant B. and Elmasri Ramez. 2000, Data Warehousing and Data Mining, in 'Fundamentals of Database Systems'. Pearson Education pvt Inc. Singapore, 841-872.

[24] Lizhen Liu and Junjie Chen 2002. The Research of Web Mining, *Proceedings of the 4th World Congress on Intelligent Control and Automation* June 10-14, 2002, Shanghai, P.R. China, IEEE. 2333-2337.

[25] Brin S., and Page L.1998. The anatomy of a largescale hyper textual Web search engine, Computer Networks and ISDN Systems, 30(1-7): 107-117.

[26] Kleinberg J.M., 1999, Authoritative sources in hyperlinked environment, Journal of ACM, Vol.46, No.5, 604-632.

[27] JISC, 2008. Text Mining Briefing Paper, *Joint Information Systems Committee*, accessed from http://jisc.ac.uk/media/documents/publications/bptextminingv2.pdf (27 October 2009).

[28] Dahl Stephan. 2010. 'Current Themes in Social Marketing Research: Text-Mining the Past Five Years', Social Marketing Quarterly, 16: 2, and 128-136.

[29] Kim J., Ohta T., Tsuruoka Y., Tateisi Y. and Collier N. 2004. Introduction to the Bio-Entity task at jnlpba. In N. Collier, P. Ruch, and A. Nazarenko, editors, *Proc. Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 70-76, 2004.

[30] PaaB G. and deVries H. 2005. Evaluating the performance of text mining systems on real-world press archives. In Proc. 29th Annual Conference of the German Classification Society (GfKl 2005), pages 61-70, Oulu, Finland, Sep 2001. Infotech.

[31] Kaushik, Abhishek and Naithani, Sudhanshu 2016. A Comprehensive Study of Text Mining Approach. *International Journal of Computer Science and Network Security*, Vol. 16 No. 2, February 2016. http://paper.ijcsns.org/07_book/201602/20160212.pdf.

[32] Fan W., Gordon M. D., and Pathak P. 2006. Personalization of search engine services for effective retrieval and knowledge management. *In Proceedings of the 21th International Conference on Information Systems*, pages 20-34, 2006.

[33] Stavrianou Anna, Andritsos Periklis and Nicoloyannis Nicolas. 2007. Overview and Semantic Issues of Text Mining. SIGMOD Record, September 2007. Vol. 36, No.3.

[34] McCallum Andrew. 2005. Information Extraction: Distilling Structured Data from Unstructured Text. ACM Queue, 3(9), November 2005.

[35] Metzler, D., Bernstein, Y., Croft, W. B., Moffat, A., and Zobel, J. 2005. Similarity measures for tracking information flow. In Proc. of CIKM, Bremen. Germany, pp. 517-524.

[36] Nenadic, G., and Ananiadou, S. 2006. Mining semantically related terms from biomedical literature. In ACM TALIP Special Issue on Text Mining and Management in Biomedicine, 5(1), pp. 22-43.

[37] Adeyemo A.B. and Ojo A.K. 2014, "Classification of Social Blogs Comments Using Text Mining". International Journal of Computer Science Issues (IJCSI), Vol. 11 No. 6. Pp. 54-58 www.IJCSI.org.