Knowledge Discovery In Academic Electronic Resources Using Text Mining

Ojo, Adebola K.
Department of Computer Science
University of Ibadan
Ibadan, Nigeria

Adeyemo, Adesesan B.
Department of Computer Science
University of Ibadan
Ibadan, Nigeria

Abstract - Academic resources documents contain important knowledge and research results. They have highly quality information. However, they are lengthy and have much noisy results such that it takes a lot of human efforts to analyse. Text mining could be used to analyse these textual documents and extract useful information from large amount of documents quickly and automatically. In this paper, abstracts of electronic publications from African Journal of Computing and ICTs, an IEEE Nigerian Computer Chapter Publication were analysed using text mining techniques. A text mining model was developed and was used to analyse the abstracts collected. The texts were transformed into structured data in frequency form, cleaned up and the documents split into series of word features (adjectives, verbs, adverbs, nouns) and the necessary words were extracted from the documents. The corpus collected had 1637 words. The word features were then analysed by classifying and clustering them. The text mining model developed is capable of mining texts from academic electronic resources thereby identifying the weak and strong issues in those publications.

Keywords: Text Mining, Academic Journals, Classification, Clustering, Document collection.

1. INTRODUCTION

Text Mining is a process of extracting new, valid, and actionable knowledge dispersed throughout text documents and utilizing this knowledge to better organize information for future reference. Mining implies extracting precious nuggets of ore from otherwise worthless rock [1]. It is the gold hidden in mountains of textual data [2].

Text mining, otherwise known as Text Data Mining (TDM), is the discovery by computer of new, previously unknown information, by automatically extracting information from a usually large amount of different unstructured textual resources. *Previously unknown* implies discovering genuinely new information. *Unstructured* means free naturally occurring texts- as opposed to HyperText Markup Language (HTML), eXtensible Markup Language (XML), and other scripting languages.

Text mining can be described as data mining applied to textual data. Text is ―unstructured, amorphous, and difficult to deal with‖ but also ―the most common vehicle for formal exchange of information.‖ [3].

1.1 TDM and Information Retrieval

TDM is a non-traditional information retrieval (IR) whose goal is to reduce the effort required of users to obtain useful information from large computerized text data sources. Traditional IR often simultaneously retrieves both ―too little‖ information and ―too much‖ text [4] [3]. However, in Information Retrieval (Information Access), no genuinely new information is found. The desired information merely coexists with other valid pieces of information.

1.2 TDM, Computational Linguistics and Natural Language Processing (NLP)

If we extrapolate from data mining on numerical data to data mining from text collections, it is discovered that there already exists a field engaged in text data mining: corpus-based computational linguistics! Computational linguistics refers to the long-established interdisciplinary field at the intersection of linguistics, phonetics, computer science, cognitive science, artificial intelligence and formal logic, which again is frequently assisted by statistical techniques [5] [6]. Empirical computational linguistics computes statistics over large text collections in order to discover useful patterns. These patterns are used to inform algorithms for various sub problems within natural language processing, such as part-of-speech tagging and word sense disambiguation [1].

NLP is the branch of linguistics which deals with computational models of language. NLP has several levels of analysis: phonological (speech), morphological (word structure), syntactic (grammar), semantic (meaning of multiword structures, especially sentences), pragmatic (sentence interpretation), discourse (meaning of multi-sentence structures), and world (how general knowledge affects language usage) [7]. When applied to IR, NLP could in principle combine the computational (Boolean, vector space, and probabilistic) models‘ practicality with the cognitive model‘s willingness to

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

wrestle with *meaning*. NLP can differentiate *how* words are used such as by sentence parsing and part-of-speech tagging, and thereby might add discriminatory power to statistical text analysis. [3].

1.3 TDM and Data Mining (DM)

In Text Mining, patterns are extracted from natural language text rather than databases. The input is free unstructured text, whilst web sources are structured. Table 2 presents a summarized comparison of Data Mining and Text Data Mining.

Table 2: A Comparison of Data Mining and Text Mining

	DM	TM
Object of Investigation	Numerical and categorical data	Textual Data
Object structure	Structured (Relational database)	Unstructured or Semi-structured (Free form texts)
Goal	Predict outcomes of future situations	Retrieve relevant information, distil the meaning, categorize and target-deliver
Methods	Machine learning: SKAT, DT, NN, GA	Indexing, special neural network processing, linguistics, ontologies
Current market size	100,000 analysts at large and midsize companies	100,000,000 corporate workers and individual users
Maturity	Broad implementation since 1994	Broad implementation starting 2000

The relationship of data mining, information retrieval, statistics, web mining, computational linguistics and natural language processing, to text data mining is shown in Figure 2.

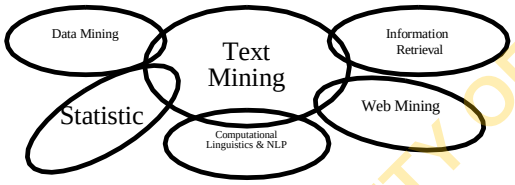


Figure 2: Relationship of Text Mining and Other Applications

2. RELATED WORK

The evolution of internet as a means for sending information led to the growth of on-line knowledge resources and to the diversification of forms and formats used for their storage and transmission: text, data, video and audio. Although hardware restrictions of storage space and data transmission speed is no longer a problem, the text still remains the most efficient form for presenting knowledge over the internet, compared to different audio, video and multimedia formats [8].

With the rapid development of the Internet, the volume of semi-structured and unstructured textual data such as XML documents, e-mail messages, blog posts, academic papers has been under an exponential growth. Discovering useful knowledge from such huge volume of data has become a very challenging problem. Text mining tries to extract knowledge from unstructured data by using techniques from data mining, machine learning, natural language processing, information retrieval, and knowledge management [9]. Text mining is a knowledge-intensive process in which a user interacts with a document collection by a suit of analysis tools, and finally identifies and explores some interesting patterns [9]. Text data mining is a natural extension of data mining [1], and follows steps similar to those in DM. The qualitative difference in text mining, however, is that TDM 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[11]. The software converts the unstructured data formats of articles, complaint forums, or Web pages into topic structures and semantic networks which are important data drilling tools. 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]. [13] 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.

[14] classifies text mining techniques into classifier learning, clustering, and topic identification. Classifiers for documents are useful for many applications. Major uses for binary classifiers include spam detection and personalization of streams of news articles. Multiclass classifiers are useful for routing messages to recipients. Most classifiers for documents are designed to categorize according to subject matter. However, it is also possible to learn to categorize according to qualitative criteria such as helpfulness for product reviews submitted by consumers. In many applications of multiclass classification, a single document can belong to more than one category, so it is correct to predict more than one label. This task is specifically called multi-label classification. In standard multiclass classification, the classes are mutually exclusive, that is, a special type of negative correlation is fixed in advance. In multi-label classification, it is important to learn the positive and negative correlations between classes [14]. Another way to view text data mining is as a process of exploratory data analysis that leads to heretofore unknown information, or to answers for questions for which the answer is not currently known. [1]

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

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 analyzed 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 [[15] [16].

2.1 Knowledge Management

There is no universally accepted definition of exactly what knowledge is. Some authors define it as the information individuals possess in their minds. This definition is argued by saying that raw data (raw numbers and facts) exist within an organisation. After processing these data they are converted into information and, once it is actively possessed by an individual, this information in turn becomes knowledge. [17] defines knowledge as the justified belief that increases the capacity of an entity to take effective action. Knowledge management is considered as the process of converting the knowledge from the source available to an organisation and then connecting people with that

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 analyzed in order to extract core concepts. Text-mining achieves this, by applying techniques from information retrieval (such as

Applications of text mining methods are diverse and include Bioinformatics [27], Customer profile analysis, Trend analysis, Anti-Spam Filtering of Emails, Event tracks, Text Classification for News Agencies, Web Search and Patent Analysis [27].

Applications of text mining can 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 source of a *news article* [32] and monitoring inconsistencies between *databases and literature.* [33]. [34] presents the framework of the proposed work.

knowledge. The aim of knowledge management is the *creation*, *access* and *reuse* of knowledge [17].
Traditionally, textual elements are extracted and applied in the data mining phase aiming to reveal useful patterns [18]. [19] concentrated on the extraction of textual elements (that is, entities and concepts). Thus the extraction and correlation of textual elements are the basis for the data mining and information retrieval phases aiming to promote support to knowledge management applications.
Knowledge management is seen as systematic and disciplined actions in which organisation can take advantage to get some return [20]. According to [21], knowledge management is an important tool for the documents may be used in order to populate and update scientific database [29]. Other areas include updating automatically a *calendar* by extracting data from *e-mails* [30], [31], identifying the original enhancement of the organisational knowledge infrastructure. The information technology has an important role in the process of transformation of the knowledge, from ***tacit*** to ***explicit*** [22]. Thus we state making explicit entities and their relationships through information extraction and retrieval, and text mining techniques is an important step towards knowledge management applications, such as, communities of practice [23], [24], expertise location [22] and competency management [25], [26].
Text-mining is ideally suited to extract concepts out of large amounts of text for a meaningful analysis. It 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 [15] [16].

3. METHODOLOGY

The overall process of conducting text-mining-based analysis goes through several steps. This is depicted in Figure 3 below. First of all, text collection and text pre-processing are the preliminary steps.

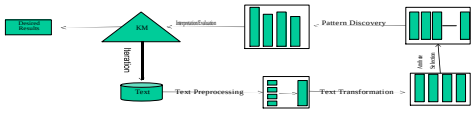


Figure 3: Text Mining Process

Second, raw journal article documents are transformed into structured data. In relation to this analysis, text mining is used as a data processing and information-extracting tool. For mining document

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

collections the text documents are pre-processed and the information stored in a data structure. A text document can be represented by a set of words, that is, a text document is described based on the set of words contained in it (bag-of-words representation). However, in order to be able to define at least the importance of a word within a given document, usually a vector representation is based, where for each word a numerical ―importance‖ value is stored.

Text Pre-processing

In order to obtain all words that are used in a given text, a *tokenization* process is required, that is, a text document is split into a stream of words by removing all punctuation marks and by replacing tabs and other non-text characters by single white spaces. This tokenized representation is then used for further processing. The set of different words obtained by merging all text documents of a collection is called the *dictionary* of a document collection.

In order to allow a more formal description of the algorithms, we define some terms and variables that will be frequently used in the following: Let D be the set of documents and $T = \{t_1, \ldots, t_m\}$ be the dictionary, that is, the set of all different terms occurring in D, then the absolute frequency of term $t \in T$ in document $d \in D$ is given by tf(d,t). We denote the term vectors $\vec{d} := (tf(d,t_1), \ldots, tf(d,t_m))$. We also need the notion of the centroid of a set X of term vectors. It is defined as the mean value $\vec{X} := \frac{1}{|X|} \sum_{d \in X} \vec{d}$ of its term vectors. In the sequel, we will apply *tf* also on subsets of terms: For $T' \subseteq T$, we let ft(d, T'):= $\sum_{t \in T'} tf(d,t)$

Text Transformation and feature selection

In order to reduce the size of the dictionary and thus the dimensionality of the description of documents within the collection, the set of words describing the documents can be reduced by filtering and lemmatization or stemming methods.

Filtering, Lemmatization and Stemming

Filtering methods remove words from the dictionary and thus from the documents. A standard filtering method is stop word filtering. The idea of stop word filtering is to remove words that bear little or no content information, like articles, conjunctions, prepositions. Furthermore, words that occur very seldom are likely to be of no particular statistical relevance and can be removed from the dictionary [27]. In order to further reduce the number of words in the dictionary, also (index) term selection methods can be used.

Lemmatization methods try to map verb forms to the infinite tense and nouns to the singular form. However, in order to achieve this, the word from has to be known, that is, the part of speech of every word in the text document has to be assigned. Since this tagging process is usually quite time consuming and still error-prone, in practice frequently stemming methods are applied.

Stemming methods try to build the basic forms of words, that is, strip the plural ‗s‘ from nouns, them ‗ing‘ from verbs, or other affixes. A stem is a natural group of words with equal (or very similar) meaning. After the stemming process, every word is represented by its stem. A well-known rule based stemming algorithm has been originally proposed by Porter (1980). He defined a set of production rules to iteratively transform (English) words into their stems.

Index Term Selection

To further decrease the number of words that should be used also indexing or keyword selection algorithms can be used. In this case, only the selected keywords are used to describe the documents. A simple method for keyword selection is to extract keywords based on their entropy. For each word t in the vocabulary the entropy can be can be computed as

$$w(t) := 1 + \frac{1}{\log_2 |D|} \sum_{d \in D} p(d,t) \log_2 p(d,t) \quad \text{with} \quad p(d,t) := \frac{tf(d,t)}{\sum_{t' \in d} tf(d,t')} \tag{2}$$

Here the entropy gives a measure how well a word is suited to separated documents by keyword search. Words that occur in many documents will have low entropy. The entropy can be used as a measure of the importance of a word in the given domain context. As index words a number of words that have a high entropy relative to their overall frequency can be chosen, that is, of words occurring equally often those with the higher entropy can be preferred.

In order to obtain a fixed number of index terms that appropriately cover the documents, a simple greedy strategy is applied: From the first document in the collection we select the term with the highest relative entropy as an index term. Then we mark this document and all other documents containing this term. From the first of the remaining unmarked documents we select again the term with the highest relative entropy as an index term. We then mark again this document and all other documents containing this term. We repeat this process until all documents are marked, and then we unmark them all and start again. The process can be terminated when the desired number of index terms has been selected.

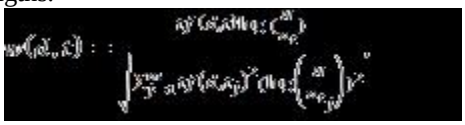
http://sites.google.com/site/ijcsis/
ISSN 1947-5500

The Vector Space Model

Despite of its simple data structure without using any explicit semantic information, the vector space model enables very efficient analysis of huge document collections.

The vector space model represents documents as vectors in *m*-dimensional space, that is, each document *d* is described by a numerical feature vector $w(d) = (x(d,t_1), \ldots, (x(d,t_m))$. Thus, documents can be compared by use of simple vector operations and even queries can be performed by encoding the query terms similar to the documents in a query vector. The query vector can then be compared to each document and a result list can be obtained by ordering the documents according to the computed similarity [27]. The main task of the vector space representation of documents is to find an appropriate encoding of the feature vector.

Each element of the vector usually represents a word (or a group of words) of the document collection, that is, the size of the vector is defined by the number of words (or groups of words) of the complete document collection. The simplest way of document chances of being retrieved independent of their lengths:



(3)

Where N is the size of the document collection D and n_t is the number of documents in D that contain term *t*.

Based on a weighting scheme a document *d* is defined by a vector of term weights $w(d)=_{(w(d,t1)}, \ldots,$ A frequently used distance measure is the Euclidian distance. We calculate the distance between two text documents d_1, d_2 as follows:

(5)

However, the Euclidean distance should only be used for normalized vectors, since otherwise the different lengths of documents can result in a smaller distance between documents that share less words than between documents that have more words in common and should be considered therefore as more similar. For normalized vectors the scalar product is not much different in behaviour from the Euclidean distance, since for two vectors and it is

(6)

encoding is to use binary term vectors, that is, a vector element is set to one of the corresponding word is used in the document and to zero if the word is not. This encoding will result in a simple Boolean comparison or search if a query is encoded in a vector. Using Boolean encoding the importance of all terms for a specific query or comparison is considered as similar. To improve the performance usually term weighting schemes are used, where the weights reflect the importance of a word in a specific document of the considered collection. Large weights are assigned to terms that are used frequently in relevant documents but rarely in the whole document collection (Hotho, et al 2005). Thus a weight *w(d,t)* for a term *t* in document *d* is computed by term frequency *tf(d,t)* times inverse document frequency *idf(t)*, which describes the term specificity within the document collection. In Salton, et al (1994) a weighting scheme was proposed that has meanwhile proven its usability in practice. Besides term fre[illegible]verse document frequency - defined as  -, a length normalization factor is used to ensure that all documents have equal $(w(d,t_m))$ and the similarity *S* of two documents d_1 and d_2 (or the similarity of a document and a query vector) can be computed based on the inner product of the vectors (by which - if we assume normalized vectors - the cosine between the two document vectors is computed), that is,

(4)

Part-of-speech tagging (POS) determines the part of speech tag, for example, noun, verb and adjective for each term.

Text chunking aims at grouping adjacent words in a sentence. An example of a chunk is the noun phrase ―the current account deficit‖.

Word Sense Disambiguation (WSD) tries to resolve the ambiguity in the meaning of single words or phrases. An example is _bank' which have - among others - the senses _financial institution' or the _border of a river or lake'. Thus, instead of terms the specific meanings could be stored in the vector space representation. This leads to a bigger dictionary but considers the semantic of a term in the representation.

Parsing: This produces a full parse tree of a sentence. From the parse, we find the relation of each word in the sentence to all the others, and typically also its function in the sentence (for example, subject, object).

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

The algorithm [35] for text extraction is given as:

{

1 Convert the text into a LIST of words

2 Set threshold to a certain value such as 1 or 2, put a separator to the end of LIST and Set an array LIST[N], an array FinaList[N]=0,

3 Do

{

3.1 Set the frequency of the separator (separator=0)

3.2 Set MergerList[N]=0,

3.3 For i from 1 to NumOf(LIST) - 1 step 1

{

3.4 If LIST[i] is the separator, then Go to Label 3.3.

3.5If Freq(LIST[i]>threshold and Freq(LIST[i+1] > threshold, then

Merge LIST[i] and LIST[i+1] into MergeList

Else

If Freq(LIST[i])> threshold LIST[i] did not merge with LIST[i-1], then

Save LIST[i] into FinaList.

If the last element of MergeList is not the separator, then

Put the separator to the end of MergeList.

}

4 Set MergeList to LIST

}while NumOf(LIST) <2

5 Filter terms in FinaList

}

4. Results and Discussion

Document Collection

This involves the gathering of academic journal articles using academic electronic resources from African Journal of Computer and ICT, IEEE Nigerian Section.

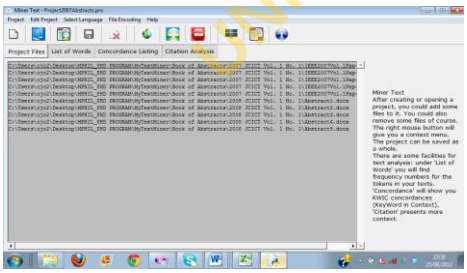


Figure 4: Document Collection

Text Extraction: This involves the identification and extraction of texts from those scientific publications. These raw article documents are then transformed into structured data as shown in Figure 5 below:

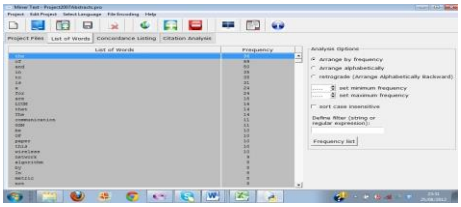


Figure 5: Text Extraction

THE CLUSTERING RESULTS: Overview of the Data

Keywords:

- Data Communication (D):

Broadcast, Radio, acoustic, transmitters, receivers (5)

- **Technology/ICT (T):**

Hardware, Software, Storage device, Coding, Computers, Electronics (7)

- **Location (L):**

world, country, Nigeria (3) •

Field/Discipline (F):

Science, Education, Engineering, Medical (4)

- **Product/Market (P):**

result, expansion, advertiser, advancement, economy, present, exploration, finances (8)

- **Organisation (O):**

Government, professionals, subscribers, entrepreneurship (4)

- **Papers/Journal (J):**

published, research, scholars, review (4)

- **Unit (U):**

Age, number, year (3)

- **Facility (Y):**

BCOS, NTA, AIT, Channel, television (5)

- **Method (M):**

Approaches, Measures, techniques, factors (4)

- **Person (N):**

Noble, group, we, I (4)

- **Miscellaneous (S):** other words which did not fall into any of the categories above.

(The numbers in the parenthesis indicate the total number of keywords used during text search.)

4.1 Text Pre-Processing, Transformation and Feature Selection

These involve Text Clean up and tokenization. The document is split into a series of words (features). Stop Words were removed, and words stemmed down to their roots.

4.2 Attribute Generation

Attributes generated are merely labels of the classes automatically produced by a classifier on the features that passed the feature selection process. After this, the database is populated as a result of the process above.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table 3: Attribute Generation

ABSTRACT	DATA COMM	TECHNOLO	LOCATIO	FIELD/DISC	PRODUCT/	ORGAN	PAPERS/	UNIT	FACILITY	METHOD	PERSON	MISCELAS	STOP WO	TOTAL
1	0	20	19	4	27	1	21	9	0	0	0	53	142	296
2	3	8	8	7	14	3	0	3	0	0	1	27	45	119
3	36	18	1	0	26	0	1	4	0	1	5	18	55	165
4	25	4	8	0	29	0	0	5	17	0	1	15	75	179
5	9	16	0	2	18	1	1	12	0	19	0	13	59	150
6	28	25	0	0	2	0	1	11	0	15	4	27	68	181
7	0	3	1	0	9	0	0	7	0	6	1	12	56	95
8	34	2	4	0	9	0	1	2	1	18	0	21	59	151
9	0	2	1	0	7	0	2	5	0	39	1	22	45	124
10	0	2	0	0	13	0	0	3	0	51	24	11	73	177
TOTAL	135	100	42	13	154	5	27	61	18	149	37	219	677	1637

From Table 3, the corpus consists of abstracts taken from the journal articles (as a sample), having a total number of 1637 words including keywords, title words, and the clue words. The rest are stop words. The keywords, title words and the clue words are all categorised as Data Communications (e.g. transmitters, receivers, bandwidth, broadcast, radio link), Technology/ICT(e.g. software, hardware, devices, computers), Location (e.g., world, Nigeria, Africa, country), Field/Discipline (e.g. Science, Education, Engineering), Product/Market (result, economy, expansion), Organisation (Government, entrepreneurship, professionals), Papers/Journals (research, review, published), Unit, Facility (age, number, year), Methods (approaches, techniques, algorithms, measures), Person (person, noble, group), and Miscellaneous (e.g. used, suggests, offers). Stop words include words such as _the‘, _is‘, _of‘, and _to‘.

Table 4:Attribute Selection

ABSTRACT	DATA COMM	TECHNOLO	LOCATIO	FIELD/DISCI	PRODUCT/	ORGANI	PAPERS/	UNIT	FACILITY	METHOD	PERSON	MISCEL	STOP WOR
1	0	4	4	1	6	1	5	2	0	0	0	3	8
2	1	2	2	2	3	1	0	1	0	0	1	2	3
3	8	4	1	0	6	0	1	1	0	1	1	1	3
4	5	1	2	0	6	0	0	1	4	0	1	1	4
5	2	4	0	1	4	1	1	3	0	4	0	1	3
6	6	5	0	0	1	0	1	3	0	3	1	2	4
7	0	1	1	0	2	0	0	2	0	2	1	1	3
8	7	1	1	0	2	0	1	1	1	4	0	2	3
9	0	1	1	0	2	0	1	1	0	8	1	2	3
10	0	1	0	0	3	0	0	1	0	11	5	1	4

Table 4 was generated from Table 3 using the following class intervals: 1 (1-5), 2 (6-10), 3(11-15), 4(16-20), 5(21-25), 6(26-30), 7(31-35), 8(36-40); and for miscellaneous data and stop words, the following class intervals: 1(1-20), 2(21-40), 3(41-60), 4(61-80), 5(81-100); 6(101-120), 7(121-140), 8(141-160), and 9(161-180). This is to reduce the population of data. By taking each attribute as an effect, Probability Models were generated from Table 4, by taking Probability [redacted]. The resulting output was given in Table 5.

Table 5: Probability of Occurrence Of Each Attribute

EVENT	DATA COMM	TECH/ICT	LOCATION	FIELD/DIS	PRO/MKT ORG		PAP/JO	UNIT	FACILITY	METHOD	PERSON	MISCE	STOP WOR
RAW DATA (N=10)	0	1	1	1	1	1	1	1	0	0	0	1	1
	1	1	1	1	1	1	0	1	0	0	1	1	1
	1	1	1	0	1	0	1	1	0	1	1	1	1
	1	1	1	0	1	0	0	1	1	0	1	1	1
	1	1	0	1	1	1	1	1	0	1	0	1	1
	1	1	0	0	1	0	1	1	0	1	1	1	1
	0	1	1	0	1	0	0	1	0	1	1	1	1
	1	1	1	0	1	0	1	1	1	1	0	1	1
	0	1	1	0	1	0	1	1	0	1	1	1	1
	0	1	0	0	1	0	0	1	0	1	1	1	1
TOTALPROB. OF	6	10	7	3	10	3	6	10	2	7	7	10	10
EOCCURRENC	0.6	1	0.7	0.3	1	0.3	0.6	1	0.2	0.7	0.7	1	1

In Table 5, each attribute is taken as an event. When an event occurs, the attribute is assigned 1; otherwise, it is assigned zero (0). It is observed from the above that probabilities of data in Groups Technology/ICT and Product/Market are one (1). This means that most of these journals concentrated on the category Technology/ICT, which involves the use of hardware, software, devices, computers and electronics.

Furthermore, it was discovered that stop words had the highest frequency in the whole corpus. After filtering, there was more concentration on Products/Market, and Methods used. This is further represented graphically in Figures 6, 7 and 8:

Figure 6: All Attributes Considered

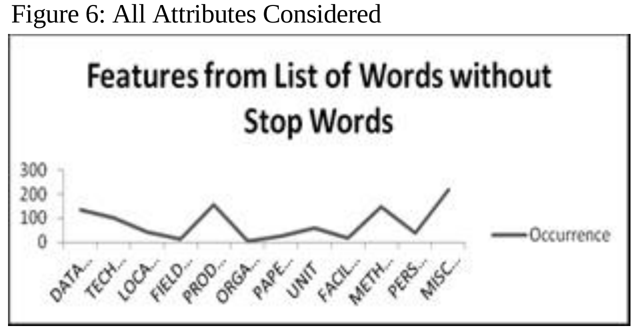


Figure 7: All Attributes Without Stop Words

Figure 8: All Attributes Without Stop Words and Miscellaneous

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table 6: Correlations among the Attributes

	Data Communication	ICT	Location	Field & Discipline	Product &Market	Organisation
Data Communication	1	.069	-.768	-1.000**	.032	-1.000**
I CT	.069	1	.546	-.737	.161	-.945
Location	-.768	.546	1	-1.000**	.551	-1.000**
Field and Discipline	-1.000**	-.737	-1.000**	1	-.408	.918
Product and Market	.032	.161	.551	-.408	1	-.737
Organisation	-1.000**	-.945	-1.000**	.918	-.737	1

Table 6 shows the correlations (relationships) among all the attributes. It was discovered that there were correlations between some attributes: between attributes ICT and Location (0.546) where ICT was the dependent variable while Location was independent; Product and Location (0.551) where Product was a dependent variable while Location was independent; Paper/Journal and Methods Used (0.847) where former was a dependent variable while the latter was the independent one.

5. Conclusion

Academic resources documents contain important knowledge and research results. They have highly quality information. However, they are lengthy and have much noisy results such that it takes a lot of human efforts for analysis. Text mining could be used to analyse these textual documents and extract useful information from large amount documents quickly and automatically.

This study provides a method for analysing unstructured text. The software captures some selected abstracts of academic publications from the universities electronic resources websites. The processed data was then ‗mined' to identify patterns and extract valuable information and new knowledge. The study revealed some strong areas of focus by the authors of these articles in this journal while less concentration was on other areas. This will enable us to have a greater understanding of the patterns and trends of data in these journal articles in future. It will be useful to shape the debate about future research and publications, and hopefully engage current authors of these articles to go beyond the most published (Data Communications) and into other areas of applications.

This research work was based on the academic resources in a particular journal which was specifically based on Data Communications. It can thus be extended to cater for all the journal articles, which cut across other disciplines and fields in Computer Science, as well as all other areas and disciplines in the academic world. This will enable us to know the trends of those publications when taken periodically. Furthermore, it can also be extended to texts being generated by business, academic and social activities - in for example competitor reports, research publications, or customer opinions on social networking sites to capture knowledge and trends.

RERERENCES

[1] M. Hearst, (1999) ―Untangling Text Data Mining,‖ in the *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*.

[2] Dorre, J., Gersl, P., & Seiffert, R. 1999. Text mining: Finding nuggets in mountains of textual data. (KDD-99, Association of Computing Machinery, 8, 223-239.

[3] Sharp, M. 2001. Text Mining. Term Paper in Information Studies. Rutgers University, School of Communication, Information and Library Studies. 11 December 2001

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

[5] Clegg, A. B. 2008. Computational-Linguistic Approaches to Biological Text Mining. A PhD Thesis submitted to the

School of Crystallography, Birkbeck, University of London, Malet Street, UK.

[6] Jurafsky, D. and Martin, J. H. 2000. Speech and Language Processing. Prentice Hall, New Jersey.

[7] Bird, S., Klein, E., and Loper, E. 2007. Natural Language Processing in Python. Draft Copy. University of Pennsylvania. October 12, 2007

[8] Vespan, Dragos M., (2009). PhD Thesis Review: Knowledge Acquisition through Text Mining. *Informatica Economică*, vol. 13, no. 2/2009

[9] L. Jing and R. Y. K. Lau (2009). Granular Computing for Text Mining: New Research Challenges and Opportunities. SpringerLink. Abstract.

[10] Kuan C. Chen, (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),

[12] Cerrito, Patricia. Inside Text Mining. March 24, 2005

[13] Louise Francis and Matt Flynn. (2010) Text Mining Handbook. Casualty Actuarial Society *E-Forum*

[14] Elkan, C. 2011. Text Mining and Topic Models. elkan@cssd.edu

[15] JISC, 2008. Text Mining Briefing Paper, Joint Information Systems Committee, accessed from

[16] Dahl, Stephan (2010) _Current Themes in Social Marketing Research: Text-Mining the Past Five Years', Social Marketing Quarterly, 16: 2, 128 — 136

[17] Nonaka, I., von Krogh, G. 2009. "Tacit Knowledge and Knowledge Conversion: Controversy and Advancement in Organizational Knowledge Creation Theory". *Organization Science* **20** (3): 635-652. doi:10.1287/orsc.1080.0412.

18] Mooney, R. J. and Nahm, Un Y. (2005) ―Text Mining with Information Extraction‖. In: Proceedings of the 4th International MIDP colloquium, September 2003, Bloemfontein, South Africa, Daelemans, W., du Plessis, T., Snyman, C. and Teck, L. (Eds.), Van Schaik Pub., South Africa, p. 141-160, 2005.

[19] Goncalves, A. L., Beppler, F. Bovo, A., Kern, V. and Pacheco, R. 2006. A Text Mining Approach Towards Knowledge Management Applications.

[20] Davenport, T. H. and Prusak, L. (1997). ―Information ecology: Mastering the information and knowledge environment‖, Oxford University Press.

[21] Schreiber, G., Akkermans, H., Anjewierden, A., Hoog, R. de, Shadbolt, N., Velde, W. V. de and Wielinga, B. (2002), Knowledge engineering and management: The CommomKADS Metlhodology, The MIT Press, 3rd edition.

[22] Marwick, A.D. (2001) ―Knowledge management technology‖. IBM Systems Journal, v. 40, n. 4, p. 814-830.

[23] Lesser, E. L. and Storck, J. (2001) ―Communities of practice and organizational performance‖, IBM Systems Journal, v. 40, n. 4, p. 831-841.

[24] Wenger E. (1998), Communities of practice, learning meaning and identity, Cambridge University Press, Cambridge, MA.

[25] Dawson, K. (1991) ―Core competency management in R&D organizations‖, In *Technology Management: The New International Language*, Dundar Kocaoglu and Kiyoshi Niwa (eds.), New York, Institute of Electrical and Electronics Engineers, p. 145-148.

[26] Hafeez, K., Zhang, Y. and Malak, N. (2002) ―Identifying core competence‖, IEEE Potencials, v.
49, n. 1, p. 2-8.

[27] Hotho, A., Nurnberger, A. and Paaß, G. 2005. A Brief Survey of Text Mining. Retrieved on April 4, 2011.

[28] Fan, W., Wallace, L., Rich, S. and Zhang, Z. 2006. Tapping the power of text mining. In Communications of the ACM 49(9), pp. 76-82.

[29] Swanson, D. R. and Smalheiser, N. R. 1997. An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artificial Intelligence 91, pp. 183 - 203.

[30] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell, (2000) ―Text Classification from Labeled and Unlabeled Documents using EM,‖ in *Machine Learning*, 2000.

[31] Stavrianou, A., Andritsos, P., and Nicoloyannis, N.2007. Overview and Semantic Issues of Text Mining. SIGMOD Record, September 2007 (Vol. 36, No. 3).

[32] 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.

[33] 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.

[34] Ojo, A. K. and Adeyemo, B. A. (2012). A Framework for Knowledge Discovery from Journal Articles Using Text Mining Techniques. IEEE Journal… Ojo, A. K. & Adeyemo, A. B. (2012): ―A Framework for Knowledge Discovery from Journal Articles Using Text Mining Techniques‖. African Journal of Computing & ICTs (An IEEE Nigeria Computer Chapter Publication) Vol. 5, No. 2, March, 2012 33-42 www.ajocict.net

[35] Liang Yanhong, Tan Runhua. A Text-Mining-Based Patient Analysis in Product Innovative Process. Hebei University of Technology

AUTHORS PROFILE

Dr. Adesesan Barnabas ADEYEMO is a Senior Lecturer at the Computer Science Department of the University of Ibadan. He obtained his PhD, M. Tech., and PGD Computer Science degrees at the Federal University of Technology, Akure. His research interests are in Data Mining, Data Warehousing & Computer Networking. He is a member of the Nigerian Computer Society and the Computer Professionals Registration Council of Nigeria. Dr Adeyemo is a Computer Systems and Network Administration Specialist with expertise in Data Analysis and Data Management.

Adebola K. OJO is a lecturer in the Department of Computer Science, University of Ibadan, Nigeria. She is a registered member of the Computer Professional of Nigeria (CPN). She had her Masters of Science Degree in Computer Science from University of Ibadan, Nigeria. Her research interests are in Digital Computer Networks, Data Mining, Text Mining and Computer Simulation. She is also into data warehouse architecture, design and data quality via data mining approach.