A Two-Stage Method for Scientific Papers Analysis

Damien Hanyurwimfura[1, 2], Bo Liao[1]
[1]College of Information Science and Engineering, Hunan University, Changsha, China
Email: hadamfr@gmail.com, boliao@yeah.net

Emmanuel Masabo[2], Gaurav Bajpai[2]
[2]College of Science and Technology, University of Rwanda, Kigali, Rwanda
Email: gb.bajpai@gmail.com, masabem@gmail.com

***Abstract*—A considerable amount of research is being conducted by many people (researchers, graduate students, professors etc) everyday. Finding information about a specific topic is one of the most time consuming activities of those people. People doing research have to search, read and analyze multiple research papers, e-books and other documents and then determine what they contain and discover knowledge from them. Many available resources are in the form of unstructured text format of long text pages which require long time to read and analyze. In this paper we propose a two-stage method for scientific paper analysis. The method uses information extraction to extract the main idea key sentences (mainly needed by the most readers) from the paper and the extracted paper's information is then organized in a structured format and grouped in different clusters according to their topics using a multi-word based clustering method. The proposed method combines different features in paper's topics extraction and uses multi-word matching feature in selection of initial centroids for clustering. The proposed method can help readers to access and analyze multiple research papers documents timely and efficiently. Conducted experiments show the effectiveness and usefulness of our proposed approach.**

***Index Terms*—text mining, information extraction, text clustering, important information, initial centroids, scientific papers.**

I. Introduction

The amount of textual data that is available for researchers and businesses to analyze is increasing at a dramatic rate [1]. Most of available information (over 80%) is stored as unstructured text, and text mining is believed to have a high commercial potential value [2]. As the amounts of data are increasingly available in the form of text, text analysis and text mining are both becoming topics of significant interest. This information is being read and analyzed by many different people for different purposes like knowledge discovery, for decision-making and knowledge management through text mining. Facing such large data sets, it is very difficult to find the desired information quickly and accurately. There is a need of automated processing methods for such amount of information available for many readers like researchers, academicians, students, professors and business analysts so that they access this information fast and accurately. It is very helpful for readers to retrieve the summary of original article instead of original article with the purpose of finding the desired information efficiently and accurately. It is also much needed to retrieve all needed information very fast in well structured and organized manner. Users need tools to compare different documents, rank the importance and relevance of the documents, or find patterns and trends across multiple documents [3]. Text mining is defined as an intelligent text analysis, text data mining, or knowledge discovery in text[2]- that is uncover previously invisible patterns in existing resources by applying principles from several fields such as computational linguistics, information retrieval, machine learning, and statistics[4]. Typically text mining tasks include information extraction, text clustering, text classification, information retrieval and text summarization. Researchers have put a lot of interest on mining data that are in the form of structured format where they assume that the information to be mined is already in the form of a relational database [5]. Unfortunately some research papers, e-books and news articles are in the form of unstructured or semi structured format which is not easy to apply data mining or knowledge discovery directly.

Text mining starts by extracting information (facts and events) from textual documents and then enables traditional Data Mining and data analysis methods to be applied. Document clustering (also known as text clustering) is one of the most important text mining methods that are developed to help users effectively navigate, summarize, and organize text documents[6].

To perform analysis, decision-making and knowledge management tasks and information systems use an increasing amount of unstructured information in the form of text [7].

Generally, readers (researchers) of scientific papers are only and mainly interested in the main idea of the research paper and do not need to read the whole paper at first. Summarization of a scientific paper by only extracting key sentences (main ideas) can help reader

doi:10.4304/jsw.9.10.2564-2573

reads many papers in a very short time and reduces storage size. In this approach, a two stage method for scientific papers analysis is proposed, it uses information extraction technique to extract the needed information by readers from scientific papers and store it in a structured format (database) as a summary of the papers for easier further processing. To solve the high dimensional problem found in text documents, only the extracted information is processed and used to cluster research papers in their topics based on their similarity using a multi word based clustering method. Fig. 1 shows the details of the proposed approach. In stage 1, information extraction method is used to extract the main idea from papers and the extracted information in form of sentences or phrases is stored in a database for further processing. In stage 2, the extracted information is clustered in different topics based on their similarity, which is means that papers which are talking the same topic are grouped in the same cluster while different topic papers are grouped in different clusters. With this new method, all research papers are stored in databases and their descriptions (Title, authors, keywords, main idea/key sentences, paper's topic and most similar reference title) can be queried. For example the reader or analyzer can get how many papers have been published in a certain field of research or she/he can know that such author has published such number of papers, etc. Fig. 1 shows number of papers in different topics after clustering.

Having the needed information in summarized, viewed, structured format with similar (related) papers grouped together in their well organized categories or topics will provide to the reader or analyzer an easier access by quickly accessing many papers' information, know their contents in a very short time.

Text summarization is needed because it presents information in shorter way, it saves reader's time and it reduces storage space [8].

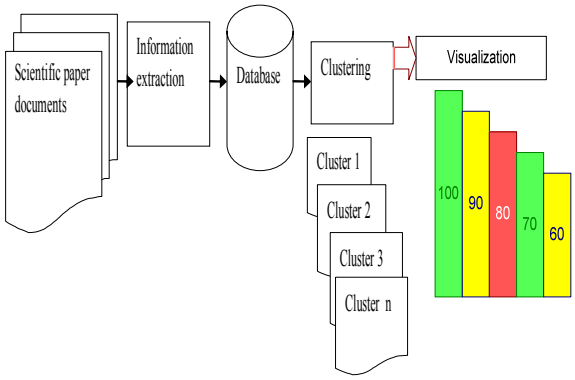

Figure 1. The proposed approach architecture.

II. Related Works

A starting point for computers to analyze unstructured text is to use information extraction [2].

Many information extraction methods have been developed for all kinds of text data. Those methods can only extract simple words from an unstructured text. Mostly useful information such as names of people, places or organization mentioned in the text is extracted without a proper understanding of the text [9]. Reference [10] proposed a scientific paper summarization using citation summary networks. In this method, it is assumed that a set of citation summary can be good resource to understand contributions of a paper and how that paper affects others. A citation summary can be a good resource to make a summary of a target paper. They used citation summaries and network analysis techniques to produce a summary of a single scientific article as a framework for future research on topic summarization.

Text clustering has been used in many applications such as text summarization [11-13], navigation of large document collections and organization of Web search results. In their experiments on identifying and clustering similar sentences from one or multiple documents of news articles, Reference [14] proposed an evaluation framework based on an incremental and unsupervised clustering method which is combined with statistical similarity metrics to measure the semantic distance between sentences. Their approach detects and clusters similar sentences of texts written in Brazilian Portuguese. Their approach is limited on working only for text written in Brazilian Portuguese and also works for only news articles. Few works have been conducted to analyze scientific papers, one of them is the analysis of scientific papers in the field of radiology and medical imaging included in Science Citation Index Expanded and published by Turkish authors [15]. The analysis was done through web science software by analyzing the number of publications per year, journals, institution and type of papers. In this study there is no clear proposed method or algorithm that can handle the automated analysis of papers and a lot of things have done manually.

III. Motivation for the Proposed Approach

Researchers spend a lot time on searching and reading papers related to their research topics. People doing research have to read many papers so that they can see what others have done and they can contribute to the existing approaches or can propose new ones. They can conduct some analysis in the interest to know the progress of research in a certain area, for example the researcher might want to know which research topics are more active or which authors has contributed to a certain area of research, etc. Given the high number of researches and the increasing amount of information on the web, it becomes very important to organize this large amount of information into meaningful clusters referring each to one single category so that the time to access them is reduced.

IV. A Two Stage For Scientific Paper Analysis

The readers of a scientific paper can be only interested in the main idea of the paper and he/she doesn't need to read the whole paper. For most papers, abstract is believed to be a good summary of the papers and can contains the results of experiments but it is not the case for all papers. An abstract is a brief summary of a research article, thesis, review, conference proceeding or

any in-depth analysis of a particular subject or discipline, and is often used to help the reader quickly ascertain the paper's purpose [16]. Mostly the main idea of the paper is found in abstract section of the paper which indicates that the abstract is the accurate summary of the content of the paper and it is often the only component which is read by many readers. Our objective is to extract important information to the reader from the paper and key sentences (topic of the paper) to help in achieving better results in clustering.

Our approach is mainly executed in two stages: Information Extraction and Text Clustering. Information extraction can help to automatically extract only the needed information to the reader and text clustering can help to group similar papers together for the benefit of the reader or analyzer. The output of clustering method shows papers in different topics and can show also graphically the number of research papers from different fields or topics as shown in Fig.1, thus speed up the analysis of research papers.

A. Information Extraction Stage

During this stage, we want to extract the main information of the paper that will be taken as a summary and used in further processing (clustering). The overall idea of the paper can be expressed by the title of the paper, but the title only is not enough for the reader to know what is the content of the paper; we need to extract some other additional information to express paper content. It is in this regard that we selected other information such as abstract to help express the main content of the paper. We believe that in an academic research paper, the following information is mainly very important to the reader; it can be taken as a summary of the paper and is selected to be extracted.

- Title: this contains the title of paper
- Authors: this contains the names of all authors of the paper.
- Abstract: this contains the summary of the paper
- Keywords: this contains common used words in the paper.
- Paper topic: this contains the key points or important points of the paper.
- Most similar reference title: this contains the most similar references to the paper's title.

Since the academic research paper is semi-structured in such way that it has sections or parts, some information will be extracted from their respective sections without much processing. Here we mean paper's title, authors, abstract and keywords but other information such as paper's topics and most similar references, some processing is required in order to extract them.

The title and authors' name, abstract and keywords are automatically extracted and stored in database.

Reference [1] described his concepts and assumptions that the fundamental unit of text is a word. Words are comprised of characters, and are the basic units from which meaning is constructed. By combining a word with grammatical structure a phrase or sentence is made. Sentences are the basic units of action in text, containing information about the action of some subjects. Since phrases or sentences are considered more meaningful than individual words, a phrase match in the document is considered more meaningful than single word matches [17].

Sentences extraction can provide better results in term of further processing. In order to get meaningful clusters, we use multi word or phrase features in clustering algorithm. In reducing the dimension of text document for fast processing only little paper's important information is considered to represent the paper. A part from paper title, its abstract and its keywords, we need also paper's topics and its most similar references title in order to make an automated paper clustering. We say paper's topics because in many cases a single paper can have many topics.

A.1 Paper's Topic Extraction

Some works have been done in localizing the main idea of the document. Many of them considered sentences feature to contain document's main idea. Selecting sentence based on its location in the document was proposed [19]. This is dependent on the type of the document. For example, in technical documents, sentences in the conclusion section are ranked high, while in news articles, first few sentences are ranked higher [20].

In extraction of paper's topics we consider features such as high frequent words, most frequent sentences which are most related to the paper's title and keywords. Each paper section is processing except table, figure, equations, footnotes, header and footer information. After removing stop words from the papers, we didn't perform stemming as we assumed that stemming would miss the tense and voice information of the sentence of the paper, so we choose the original words in the sentences. In general as a researcher when you read the title of a scientific paper, abstract and its keywords, it is easily to know or guess which kind of topics the paper is talking about.

Compared to the length of the research paper, sometimes up to 40 pages, this information is not enough to know the topics of a given paper. We have to look for other information from the remaining parts which are enough to automatically group similar papers.

We treated each research paper as a bag of words and used TF (Term Frequency) to get the frequency of each word. Only top frequent words at a certain occurrence are considered and sentences containing those top frequent words are extracted to be included in paper's topics selection. Top frequent sentences extraction algorithm is shown in Fig.2.

Input:
Top frequent words Fw= {fw_1, fw_2 …fn)
Keywords Kn = { k_1, k_2…….k_n)
All sentences of the paper= Sn= {s_1, S_2…….s_n)
Output:
Selected T sentences
steps
Kt=Fw $\bigcup$ Kn
For all Ft_i
For all Sn_j
If Sn_j contains Ft_i
Extract Sn_j
Repeat
End For
End for

Figure 2. Top Frequent Sentences extraction algorithm

In order to reduce the number of sentences and keep the title relationship, the similarity between title and each extracted sentences is measured and only those most similar the title (at a certain occurrence) are extracted.

We have chosen to use these sentences instead of single words because a multi word is meaningful than single word and we want to have meaningful paper topics. Among picked sentences only sentences very similar to the title are selected. We use Jaccard similarity to measure the closeness of the sentences to the title. That is if T is the title and S is a sentence selected then

$$Sim(T,S) = \arg\max(|(T \cap S)|/|(T \cup S)|)$$

In the same way, the similarity between paper's title and each reference title is computed and only most similar references are selected. That is if T is the title and R is a reference then

$$Sim(T,R) = \arg\max(|(T \cap R)|/|(T \cup R)|)$$

With assumption that it is unusual for the phrase which is not about the topic of a document to repeat more than two times in the document, phrase feature is used to determine the topics in which papers are talking about. In selecting paper's topics, we need paper's title, keywords, extracted top frequent sentences and most similar reference titles. Among them only paper' keywords are in the form of phrases as they have formulated by authors. The next step now is to formulate phrases from extracted paper' title, extracted top sentences and most similar references which will be used in selection of topics.After removing the stop words from paper's title, top frequent sentences and most similar references, the following regular expression is used to formulate phrases:

$((A \mid N)^+ (A \mid N)^* (A \mid N)$

Where A is an adjective, N is a noun.

Each paper's title, extracted sentences or most similar reference title is viewed as a sequence of words, so that it can be represented as = {w_1, w_2 ,w_3 ,...), where w_1 ,w_2 ,w_3 .. are words appearing in each one of them. An ordered sequence of two or more words is called a phrase. Now we have phrase from paper's title, keywords, top frequent sentences and most similar reference titles. The next step is to calculate the frequency of each phrase. The frequency of each selected phrase is computed and only top frequent phrases at a certain occurrence are selected as paper's topics candidate. Usually, short multi-words refer to the general concepts in the documents and long multi-word is a subtopic of these general topics [18]. In our method short multi-words are selected as paper's topics. Actually, there are some overlapping among the extracted multi-words where some short multi words are subset of long multi words, for example if "*sentence similarity measure*" and "*sentences similarity*" were extracted as top phrases similar to the title, we will consider only "*sentences similarity*" to be used for topics selection which means that short phrases are preferred than long phrases and are selected as paper's topics.

By extracting information from a corpus of such textual scientific research papers, a structured, searchable database of papers can be automatically constructed and can be queried, mined, etc. Table1 shows the sample of extracted information from the paper Reference [21] that is processed and stored in a database (structured format) for easy access, analysis and further processing. It shows the papers 'information in a summarized format which makes it easy to apply other text mining tasks.

TABLE I.
SAMPLE EXTRACTED INFORMATION

title	authors	abstract	keywords	Paper topics	Most similar reference title
Sentence clustering using parts-of-speech	Richard Khoury	Clustering algorithms are used in many Natural Language Processing (NLP) tasks. They have proven to be popular and effective tools to use to discover groups…	natural language processing, Part-of-speech, clustering	natural language processing, Part-of-speech, question type classification	Automatic clustering of part-of-speech for vocabulary divided PLSA language model, Keyword extraction rules based on a part-of-speech hierarchy.

B. Paper Clustering Stage

The clustering of text documents is a central technique in text mining which can be defined as grouping documents into clusters according to their topics or main contents [22]. A cluster is defined as a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. In Reference [6], the problem of document clustering is defined as follows: given a set of documents, they can be

automatically grouped into a predetermined number of clusters, such that the documents assigned to each cluster are more similar to each other than the documents assigned to different clusters. In other words, the documents in one cluster share the same topic, and the documents in different clusters represent different topics. This automatic analysis can be useful in many tasks. It can provide an overview of the contents of a large documents collection. Another benefit of clustering [2] is that documents can appear in multiple subtopics, thus ensuring that a useful document will not be omitted from search results. We apply data mining technique (clustering) on the constructed database to group similar (or related) papers together. This technique is applied to a multi word database as it is made of phrases extracted from text documents (research papers). In our case we are concerned with scientific papers and a cluster refers to a group of related papers (the papers are similar if they talk similar field or topic).

B. 1 Similarity measure and selection of initial centroids

The most important factor in a clustering algorithm is the similarity measure. All clustering algorithms are based on similarity measures and each clustering method use a similarity function. Before clustering, a similarity/distance measure must be determined [23]. Many methods have been proposed to measure the similarity between words, sentences paragraphs and documents. Document similarity is often represented by the Vector Space Model (VSM) where documents are represented by the bag of words, and the meanings of documents are presented by vectors. A well-known similarity measure is the cosine function, which is widely used in document clustering algorithms and is reported performing very well [24]. Sentences are represented by a vector of weights while computing cosine similarity. The cosine function can be used in the family of k-means algorithms to assign each document to a cluster with the most similar cluster centroid in an effort to maximize the intra-cluster similarity. In order to achieve high efficiency of our method, we have also chosen cosine similarity as it has reported performing well.

In many previous works, K-means has been used for many methods and has reported to perform well [6, 13, 23]. The k-means algorithm is based on the idea that a centroid can represent a cluster. The k-means algorithm starts with initial cluster centroids, and sentences are assigned to the clusters iteratively in order to minimize or maximize the value of the global criterion function [6]. We adopt this algorithm in grouping similar papers of our constructed database and propose a multi- word based centroid clustering method for scientific paper documents. The concept of clustering used in this work is similar to the one used in [6, 14, 23, 25]. The difference relies on estimating the initial K centroids, paper stored in structured format (database) and the amount of information used in clustering stage. K-mean has been used also in clustering of transaction databases [26].

B. 2 Selection of initial centroids

It is not easy to determine the initial centroids of clusters in a text database. Many methods have been proposed in estimating number of clusters centroids such as random selection and buckshot. The random algorithm randomly chooses *k* documents from the data set as the initial centroids [27]. We want to have a good set of initial cluster centroids in order to get better clustering results.

The strategy that we used to determine the optimal number of clusters (the number of topics) is based on the weight of frequency of selected paper's topics combined with paper's title and paper's keywords phrases available in our constructed database. We used keywords feature because they describe the most used words and are written by the authors that means that they are much related to the paper's topics. In many case keywords are written in phrase (n gram). The phrase here is defined as set of words composed of two or more words and it is supposed to be a meaningful.

The phrases are assigned weights according to their frequencies. The frequent phrases are given higher weight. If the phrase is more frequent in the corpus, it means that it appears in many papers.

Only keywords composed of two or more words are considered to be phrases. We consider only phrase because they have specific meaning than single word which has a broad meaning.

For example the word *information* is a broad word, *information extraction*; *information security*, *information sciences*, *information retrieval* etc are more specific than i*nformation*.

As we are now working with our previously constructed database, the initial centroid selection starts at first record and counts the frequency of each phrase of the title, keywords and paper's topics attributes.

After counting the frequency of each phrase, the top K frequent phrases (at certain occurrence) are selected to represent the topics in clusters, and become cluster centroids. For example if we have the following extracted information for four papers from which we want to select the centroids(underline are phrases in the paper's titles and keywords)

Title 1: Automatic text summarization based on sentences clustering and extraction.

Keywords in Title 1*:* text summarization; similarity measure; sentences clustering; sentence extractive technique

Paper's topic: text summarization, sentences clustering, sentence similarity

Title 2: Sentence Clustering Using Parts-of-Speech

Keywords in Title 2: natural language processing, Part-of-speech, clustering

Paper's topic: natural language processing, Part-of-speech, question type classification

Title 3: A new sentence similarity measure and sentence based extractive technique for automatic text summarization

Keywords in Title 3: Similarity measure, Text mining, Sentence clustering, Summarization, Evolution algorithm, Sentence extractive technique

Paper's topic: document summarization, extractive summarization, dissimilarity measure, sentence clustering, sentence extraction and summarization methods.

Title 4: Some Experiments on Clustering Similar Sentences of Texts in Portuguese

Keywords in Title 4: Sentence Similarity, Sentence Clustering, Statistical Metrics,

Paper's topic: clustering method, document clustering, statistical similarity, clustering solution, similarity threshold and similar sentences.

After removing the stop words from paper's title and counting the frequency of each phrase in title, keywords, and paper's topics, top frequent phrases are selected as initial centroids. Table 2 shows the top phrases that are most frequent and their occurrences

TABLE 2.

TOP PHRASES AND THEIR FREQUENCIES

TOPIC CANDIDATE	FREQUENCIES
sentences clustering	**7**
extractive technique	**4**
text summarization	**4**
similarity measure	**4**
measure sentences	3
sentence extractive	2
clustering summarization	2
algorithm sentence	2
evolution algorithm	2
clustering sentences	2
processing part	2
sentence similarity	2
automatic text	2
natural language	2
language processing	2
summarization similarity	2

The top four phrases (in bold) are selected as initial centroids for those papers titles. In order to maintain the cluster definition as stated in previous section (cluster similarity and dissimilarity), we measure the similarity between selected topics and if two topics are similar at certain percentage, they are combined as one topic, which means that similar topics are kept in the same cluster while dissimilar topics are kept in other clusters.

Fig.3 shows the proposed initial centroids selection algorithm:

Input: Database of *n* record
Output: Database of n records and *K* centroids clusters

Step1. Count the frequency of every phrase in title, Keywords and paper's topic attributes.
Step2. Order them by their frequency
Step3. Select phrases whose phrases have many frequencies.
Steps 4: Calculate similarity between each of the Selected phrases
Steps5: Combine similar topics as one topic
Step5. Selected phrases are used as initial centroids

Figure 2: Initial centroids selection algorithm

B. 3 Title Clustering

After selecting K initial centroids (records from database), each title is assigned to a cluster based on a distance measure (between its paper's title, keywords, and paper's topics with each of the k centroids), then k centroids are recalculated. This step is repeated until all titles are assigned to clusters. Cosine similarity measure is used to calculate similarity between them. Figure 4 shows the proposed clustering algorithm

Input: Database of *n* record and *K* centroids clusters
Output: Database of n records grouped in different K clusters according to their topics.

Step1. Select title, keywords, paper's topics and most reference titles attributes of each *k* records (centroids), records that have been selected as initial centroids
Step2. Select title, keywords, paper's topics and most similar references attributes of one record *(r)* in remaining records
Step3. Compute cosine similarities between r and *k* centroids
Step4. Put *r* in the closest cluster and recomputed the centroids.
Step5. Repeat Steps 2 to 4 until all records are processed

Figure 4. Proposed clustering algorithm

The phrase based similarity between paper's extracted information work better and faster because it deals with phrases (multi words) rather than single words.

V. EVALUATION OF THE PROPOSED APPROACH

The accuracy of the proposed clustering solution has been evaluated by using the standard Precision, Recall, Entropy and Purity metrics which have been used in previous methods [14, 21, 28, 29] as external quality measures. They measure how good the clusters are when compared with reference clusters (often manually classified clusters). The proposed approach is only useful if it is accurate and performs as expected. Therefore it is important to measure the accuracy of the new approach

on independent test data. The primary question we address in the experiments of this section is whether the automatically extracted data is relatively reliable compared to the manually constructed database.

A. Evaluation of Information Extraction

We performed our experiments on 200 scientific research papers from 5 different fields such as: data mining, wireless networking, computer architecture, image processing and information security, downloaded from internet from different journals and conferences in which we performed manual topics and most similar reference title extraction. Stop words removal was first performed for all experiments. We randomly selected 20%, 40%, 60%, 80% and 100% of our collected papers and calculated the average of precision and recall for each group of papers. For top frequent sentences extraction only sentences appearing more than 3 times was considered and for paper's topics selection only phrases appearing from 5 times was considered. We got an average of 87.5% and 88.1.2% and 87.8 of precision, recall and F-measure respectively for all papers. Table 3 shows the performance of the proposed method in terms of precision, recall and F-measure metrics.

TABLE 3.
PERFORMANCE OF THE NEW APPROACH

Number of papers	Precision (%)	Recall (%)	F-Measure (%)
40 papers (20%)	90.6%	89.3%	89.9%
80 papers (40%)	88.0%	91.0%	89.5%
120papers (60%)	89.6%	89.6%	89.6%
160 papers (80%)	84.0%	85.0%	84.5%
200 papers (100%)	85.4%	85.8%	85.6%
Average	**87.5%**	**88.1%**	**87.8%**

When compared our results with [9, 30]. It was observed that our method outperforms by 6%, %5 and 5% higher than the ones of [9] in both precision and recall and F-measure respectively. It outperforms [30] by 4% and 1.8% higher in Recall and F-measure respectively. The results indicate that the improvement of our approach is based on multi–word extraction feature and frequent word features.

The results comparison is shown in Table 4.

TABLE 4.
RESULTS COMPARISON WITH OTHER METHODS

Method	Average precision	Average recall	F-measure
Reference [9]	81.0%	83.7%	82.3%
Reference [30]	87.9%	84.2%	86.0%
Proposed method	87.5%	88.1%	87.8%

B. Paper Clustering Evaluation

Given a set of labeled records (sentences) belonging to K classes, assume the clustering algorithm partitions them into C clusters. Let N be the total number of records to be clustered and n_{ij} the number of records of the class $k_i \in K$ that are present in cluster $c_j \in C$. The Precision and Recall for k_i and c_j, denoted $P(k_i,c_j)$ and $R(k_i,c_j)$, respectively are computed as follows:

$$P(k_i, c_j) = \frac{n_{ij}}{c_j} \quad (1)$$

$$R(k_i, c_j) = \frac{n_{ij}}{k_j} \quad (2)$$

and F-measure denoted F is computed as follows

$$F(k_i, c_j) = \frac{2 * R(k_i, c_j) * P(k_i, c_j))}{(R(k_i, c_j) + P(k_i, c_j))} \quad (3)$$

The overall F-measure of a clustering solution S, denoted F(S), is calculated by using the weighted sum of such maximum F-measures for all classes, accordingly. F(S) values range from 0 to 1. It is computed as follows:

$$F(S) = \sum_{k_i \in K} \frac{|k_i|}{N} \max(c_j) \in C\{F(k_i, c_j)\} \quad (4)$$

It is based on the cluster that best describes each class k_i, that is the one that maximizes $F(k_i,c_j)$ for all j.

The second metric employed is entropy. It measures how well each cluster is organized. A perfect clustering solution will be the one in which all clusters contain sentences from a single class only. In this case the entropy is zero.

The calculation of entropy is based on the class distributions in each cluster. This is exactly what is done by Precision metric. In fact, Precision represents the probability of a sentence randomly chosen from cluster c_j to belong to class k_i. Hence, the Entropy of a cluster c_j, denoted $E(c_j)$ and computed as follows:

$$E(C_j) = -\sum_{k_i} p(k_i, c_j) \log p(k_i, c_j) \quad (5)$$

The Entropy of a whole clustering solution S , denoted E(S) , is given by the sum of the individual cluster entropies weighted by the size of the cluster and is computed as follows:

$$E(S) = \sum_{c_j} \frac{|c_j|}{N} E(c_i) \quad (6)$$

E(S) values are always positive. The smaller the E(S), the better the clustering solution is.

Another metric used is purity. The purity of a cluster c_j, denoted $Pu(c_j)$, is defined by the class k_i that maximizes the precision of that cluster. It is computed as follows:

$$Pu(c_j) = \max k_i\{P(k_i, c_j)\} \quad (7)$$

The overall purity of a clustering solution, denoted $Pu(S)$, is given by a weighted sum of the individual cluster purities. Its values range from 0 to 1.

$$Pu(S) = \sum_{c_j \in C} \frac{|c_i|}{N} P(c_j) \quad (8)$$

In summary, Entropy and Purity metrics evaluate the goodness of a clustering solution, while F-measure evaluates the effectiveness of the clustering method.

Based on our proposed method that estimates initial k centroids based on frequency of phrases of paper titles, keywords, paper's topics and most similar reference, we conducted experiments to cluster similar papers using K means method and cosine as similarity measure. We conducted our experiment on 200 paper titles that were extracted and stored in database in the first stage.

We firstly removed stop words using the common stop word methods for each paper's title, and most similar references title; we assumed that there is no need of stemming since this can change the original meaning of the sentences. We run our experiments five times and calculated the average of F-measure, entropy and purity. At initial centroid selection steps, only phrases having from 4 occurrences were considered as initial centroids. The obtained results are compared with the previous work [23] where we only considered the clustering results on academics papers experiments as it similar to our test case. Table 5 shows the results of our proposed method. Figure 5 shows details of the comparison results with previous method [23]. As it is seen, our method outperforms previous work in term of entropy and purity.

TABLE 5.
PAPERS CLUSTERING RESULTS

papers range	F-measure	Entropy	Purity
40 papers (20%)	0.899	0.187	0.910
80 papers (40%)	0.895	0.227	0.880
120papers (60%)	0.896	0.213	0.900
160papers (80%)	0.845	0.341	0.850
200 papers (100%)	0.856	0.310	0.870
average	**0.878**	**0.256**	**0.882**

Figure 5 shows details of the comparison results with previous work [23]. As it is seen in figure 5, our method outperforms previous work in term of entropy and purity.

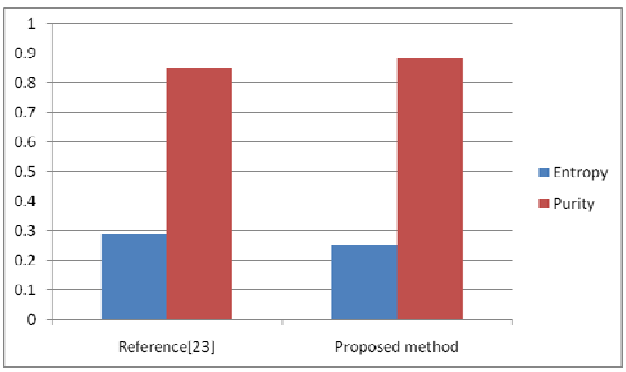

Figure 5. Clustering results comparison

In general, the smaller the Entropy value, the better the clustering result, or the larger Purity and F-measure values the better the clustering result[31].

Our clustering experiment results show an average of 87.8%, 0.256 and 0.882 of F-measure, entropy and purity respectively as shown in Table 5. The clustering results of our proposed approach give the values of 89.9%, 0.187 and 0.91 of F-measure, entropy and purity respectively for the best case. The best performance of our clustering approach is based on best extraction of topics in information extraction stage and the best selection of initial centroids while clustering by considering both papers' title, keywords, paper's topic and most similar reference title' phrases in clustering phases.

VI. CONCLUSION

A two stage method for scientific papers analysis is proposed. The method can help reader to get many papers' information and do quick analysis of research papers. It is a multi-word based clustering method for analyzing scientific papers. It extracts title, author names, abstract, keywords, paper's topics and most similar references titles from scientific research paper. The extracted information is then stored in database and processed in different clusters according to their topics and can be queried and mined easily. The main contribution of this paper is that the reader can get the paper's information in summarized, viewed, structured format for easy, and fast access, so that he or she can read the content of the paper in a very short time reducing time and space. At the same time the reader can get all related papers grouped together and this can help a reader for fast access and quick analysis of many research papers. Another contribution is that we reduced the paper's dimension by only considering a little information which resulted in fast clustering. The best performance of our clustering approach is based on the best extraction of paper's topics in information extraction stage and the best selection of initial centroids by considering both papers' title, keywords, paper's topic and most similar reference title' phrases in clustering stage. The centroids consist of phrases which are central to all papers in the same cluster, thus the closeness in papers 'topics are respected.

We are planning to improve our approach in incorporating search techniques and other topics extraction methods. This will result in excellent performance of the method in terms of accuracy and precision.

REFERENCES

[1] S. lee, J.i Song and Y. Kim, "An Empirical Comparison of Four Text Mining Methods", *J. Comput. Info. System*, Vol. 51, No. 1, pp. 1-10, 2010

[2] V. Gupta and G. S. Lehal, "Survey of Text Mining Techniques and Applications", *Journal of emerging technologies in web intelligence*, Vol. 1 No. 1, pp. 60-76, 2009

[3] J. Han and M. Kamber (edns), *Data mining Concepts and techniques*, 2nd edn. (Morgan Kaufmann. San Francisco, 2006).

[4] J. J. García Adeva and R. Calvo, "Mining Text with Pimiento", *IEEE internet computing*, pp.27-35., 2006.
[5] T. Polajnar, "Survey of Text Mining of Biomedical Corpora", pp. 1-21 ,2006.
[6] C. Luo , Y. Li and S. M. Chung, "Text document clustering based on neighbors", *J. Data & Knowl. Eng.*, 69, pp.1271–1288, 2009.
[7] A. Mustafa, A. Akbar, and A. Sultan. "Knowledge Discovery using Text Mining: A Programmable Implementation on Information Extraction and Categorization", *Int. J. of Multimedia and Ubiquitous Engineering* , Vol. 4 no. 2, pp.183-188, 2009.
[8] S. Wang, W. Li, F. Wang and H. Deng, "A Survey on Automatic Summarization". *Int. Forum on Inf. Techn. and Appl.* pp. 193-196, 2010.
[9] Y. Hu, Hang Li, Y. Cao and D. Meyerzon, "Automatic Extraction of Titles from General Documents using Machine Learning", *Information Processing and Management: an International Journal archive*, vol.42, no. 5, pp. 1276 – 1293, 2006.
[10] V. Qazvinian and D. R. Radev, "Scientific Paper Summarization Using Citation Summary Networks", In Proc. *22nd int. Conf. Computational Linguistics*. COLING ,Stroudsburg, PA, USA, pp. 689–696, 2008.
[11] F. Liu and L. Xiong, "Survey on Text Clustering Algorithm", *In Proc. int. 2nd conf. Software Engineering and Service Science* (ICSESS), IEEE, Beijing, China, pp. pp.196-199, 2011.
[12] K. Sarkar, "Sentence Clustering-based Summarization of Multiple Text Documents", *Int. J. Comput. Sci. and Commun. Tech.*, vol.2 , no. 1, pp. 225-235, 2009.
[13] Z. Pei-ying and L. Cun-he. "Automatic text summarization based on sentences clustering and extraction". *In proc. Int. 2nd Conf. Computer Science and Information Technology, ICCSIT, IEEE* , Beijing, China,pp. 167-170, 2009.
[14] E. R. Marques Seno and M. das Graças Volpe Nunes. "Some Experiments on Clustering Sentences of Texts in Portuguese". A. Teixeira et al. (Eds.): PROPOR 2008, LNAI 5190, Berlin. Heidelberg, pp. 133–142, 2008.
[15] E. Akpınar, M. Karçaaltıncaba. "Analysis of scientific papers in the field of radiology and medical imaging included in Science Citation Index Expanded and published by Turkish authors", *Diagn Interv Radiol* vol. 16. pp. 175–178, 2010.
[16] G. Blake. and W. Bly, " The Elements of Technical Writing". New York: Macmillan Publishers, 1993, pg. 117, 1993
[17] K. A. Vidhya. and G. Aghila, "Text Mining Process,Techniques and Tools: an Overview", *Int J. Inf. Tech. and Knowledge Management*, vol. 2, no. 2 , pp. 613-622, 2010
[18] Z. Wen,, Y. Taketoshi and T. Xijin " Text classification based on multi-word with support vector machine" *Knowledge-Based Systems*, vol. 21, pp. 879–886, 2008
[19] J. Goldstein, M. Kantrowitz, V. Mittal and J. Carbonell, "Summarizing Text Documents: Sentence Selection and Evaluation Metrics", in *ACM-SIGIR'99*, pp.121-128, 1999
[20] K. Madhavi and Ganapathiraju, "Relevance of Cluster size in MMR based Summarizer: A Report", 2002.
[21] R. Khoury, "Sentence Clustering Using Parts-of-Speech", *Int. J. Inf. Eng. and Electronic Business*, vol.1, pp. 1-9, 2012
[22] M. R. Aliguliyev, "A new sentence similarity measure and sentence based extractive technique for automatic text summarization", *Expert Systems with Applications*, vol. 36, no. 4, pp. 7764–7772, 2009.
[23] A. Huang, "Similarity Measures for Text Document Clustering", *NZCRSC*, pp.49- 56, 2008.
[24] M. Steinbach, G. Karypis, V. Kumar, "A comparison of document clustering techniques", *in KDD Workshop on Text Mining*, 2000.
[25] W. J. Zhang, Y. Sun, H. Wang and Y. He, "Calculating Statistical Similarity between Sentences", *J. of Conv. Inf. Tech.*, vol.6, no. 2, pp. 22-34, 2011.
[26] D. Yuan, Y. Cuan and Y. Liu, "An Effective Clustering Algorithm for Transaction Databases Based on K-Mean", *Journal of Computers*, vol.9, no. 4, pp. 812-816, 2014
[27] A. K. Jain , R.C. Dubes, *Algorithms for Clustering Data* Prentice Hall, Englewood Cliffs, 1988.
[28] C. Huang, J. Yin and F, Hou, "Text Clustering Using a Suffix Tree Similarity Measure" *Journal of Computers*, vol.6, no.10, pp. 2180-2186, 2011
[29] Y. Li, C. Luo, and S. M Chung, "Text clustering with feature selection by using statistical data", *IEEE Transactions on knowledge and Data Engineering*, vol.20 pp. 641–652, 2008
[30] J. Leskovec, N. Milic-Frayling and M. Grobelnik. "Extracting Summary Sentences Based on the Document Semantic Graph". *Microsoft Research Technical Report MSR-TR-2005-07*, January 2005
[31] L. Jing, K. Michael Ng , J. Z. Huang. "Knowledge-based vector space model for text clustering". *Knowl Inf Syst*, vol. 25, pp. 35–55, 2010.

D amien Hanyurwimfura received his Masters degree of engineering in computer science and technology from Hunan University, Changsha, China in 2010. He is currently a PhD student at the College of Information Science and Engineering, Hunan University, China. He is also a Lecturer at the College of Science and Technology, University of Rwanda, Rwanda. His current research interests include information security and data mining.

Bo Liao received the PhD degree in computational mathematics from the Dalian University of Technology, China, in 2004. He is currently a Professor at Hunan University. He was at the Graduate University of Chinese Academy of Sciences as a postdoctorate from 2004 to 2006. His current research interests include bioinformatics, data mining and machine learning.

Emmanuel Masabo has got his Master of engineering Degree in computer application technology from the School of Information Science and Engineering at Central South University, Changsha, China in 2009. He is currently a Lecturer at the College of Science and Technology, University of Rwanda, Kigali, Rwanda. His research interests include Image processing and web security.

Gaurav Bajpai received the B.Tech. degree in computer science & engineering from Rohilkhand University, India in 2000, M.Tech. degree in Software Engineering from Motilal Nehru National Institute of Technology, Allahabad, India in 2005 and Ph.D. degree in Computer Engineering from Uttar Pradesh Technical University, Lucknow, India in 2006. Currently, he is Head of Computer Engineering and Information Technology in Faculty of Engineering at KIST now referred as College of Science and Technology, University of Rwanda.

Gaurav was an Assistant Professor in the Department of Computer Science and Business Administration, Academy of Medical Sciences and Technology, Khartoum, Sudan from April 2006 to March 2007. Since March 2007, he is working as a Senior Lecturer in the Department of Computer Engineering and Information Technology, Faculty of Engineering, Kigali Institute of Science and Technology, Rwanda.

His research interests include Software Engineering, Network Routing, Network Hardware Security and Bio-Medical Engineering. He has published more than 50+ International Journal and Conference papers. He has Convened, Reviewed, been Editor and attended and presented in several workshops and seminars during his 14 year career from 2000. He has been on several International Projects with Income generation to University as well.

Dr. Bajpai is member of several distinguished organisations like ISOC, IEEE, Institution of Engineers, Life member Member ISTE,CSI etc