International Journal of Multimedia and Ubiquitous Engineering
Vol.9, No.3 (2014), pp.219-234
http://dx.doi.org/10.14257/ijmue.2014.9.3.21

A Centroid and Relationship based Clustering for Organizing Research Papers

Damien Hanyurwimfura[1,3], Liao Bo[1], Dennis Njagi[2] and Jean Paul Dukuzumuremyi[2]

[1]*College of Information Science and Engineering, Hunan University, China*
[2]*College of Information Science and Engineering, Central South University, China*
[3]*College of Science and Technology, University of Rwanda, Kigali, Rwanda*

hadamfr@yahoo.fr

Abstract

Finding research papers about particular topic of study is the most time consuming activity for many people including students, professors and researchers. 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. Huge available resources are in the form of unstructured texts format of long text pages which require a long time to process, search, read and analyze. Organizing research papers in their respective subjects or topics can facilitate the search process. We propose a new method to research paper organization and retrieval that is amenable to closely research papers and intertwined research topics. With our centroid and relationship based clustering approach, research papers are arranged and grouped within the most probable research topics or subjects. To determine topic membership, the proposed approach considers relationships such as common terms in paper title, in keywords, in referenced titles and common terms in the top frequent sentences. To solve the high dimensional problem associated with text document, only most important information of the paper is considered and we leverage on multi-word and frequent occurring phrases as the features in clustering process. Conducted experiments show that our approach is effective.

Keywords: *centroids selection, research papers, paper clustering, paper relationship, multi words features, important information*

1. Introduction

A considerable amount of research is being conducted by many people (researchers, students, professors etc) everyday. Finding information about a specific topic consumes appreciable amount of time due to the high volume of available resources. 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 referenced by distinct categories. Faced with such large data sets, it is very difficult to find the desired information quickly and accurately. Therefore, there is a need to avail automated processing methods for such large amount of information so that it is accessed fast and accurately. Text mining is defined as an intelligent text analysis, text data mining, or knowledge discovery in the text [2].

Typically text mining tasks include information extraction and text clustering [7], 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

ISSN: 1975-0080 IJMUE

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 format which is not easy to apply data mining or knowledge discovery directly. It needs other processing techniques to allow data mining to be applied. A large amount of data available on Internet is in the form of unstructured long text documents. 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. Text Mining as the automatic discovery of new, previously unknown useful knowledge from unstructured text 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].

The readers of scientific papers are usually spending a long time on the Internet in searching of papers and sometime are bored because the information they are looking for is not retrieved efficiently due to the fact that the papers are not grouped in their topics for easy and fast access. For example for some journals, research papers are not arranged in their respective topics but only grouped by volumes and numbers which make it difficult to get related papers. In this paper we propose a centroid and relationship based research paper clustering in which we group similar papers based on their relationship, which means that papers that are similar (based on the same topic) are grouped together. The papers are grouped and managed for the benefit of the reader in order to get them efficiently and easily. The method tries to solve the high dimensional problem of text document by reducing the research paper documents into small size for fast processing.

As clustering is especially useful for organizing documents to improve retrieval and support browsing [3], we can use it to group research papers. Because we want to have meaningful cluster topics, we use multi word features in clustering process in such way that initial centroids are made of phrases. Another feature in our method is that we do not need to process full text of paper, since it is a time consuming work, we only need a little important information of the paper which is enough to represent the paper's information. In automatic techniques for subject classification of research paper documents, a simple approach is to do a keyword based search for subject term or some of its synonyms in paper's title, keywords, and full text. On one hand, title and keywords provide only limited information which may lead to inaccurate decision and on the other hand, processing the whole text of a paper also takes a long time [9].

Our goal is to consider paper's little information which is relevant enough to provide accurate decisions. We believe that research papers can be related in term of common term in their titles, in common keywords; in common frequent sentences and common referenced titles (cited papers). This information can be a good resource to group similar research papers together. In this paper we want to enrich the paper's title information by considering other information much related to the paper's title instead of considering title and keywords only or the whole full text paper. Some previous proposed algorithms on text documents treat each document as a bag of words [19]. In practice, the bag-of-words model is only effective for discovering the relatedness between documents when these documents share a large proportion of lexically equivalent terms [31].

In our method, we only use the bag of words method to get frequent words which will help us to get frequent sentences. Subsequently, we treat research documents as

sequence of words with associated meaning [16]. Grouping research papers in their respective topics or subjects will help the reader to retrieve them easily and efficiently.

2. Related works

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]. Luo *et al*. [6] defined the problem of document clustering 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. 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. Mohsen [9] introduced a novel supervised approach for subject classification of scientific articles based on analysis of their interrelationships. He exploits links such as citations, common authors, and common references to assign subject to papers. In his method, he first builds a graph of relationships in which nodes represent papers and links represent the relations such as citations, common author, and common reference. Then a new algorithm based on electrical conductance to see for a given subject, how much each node in graph is relevant to that subjected was proposed. His algorithm measures the quantity of flow each node receives from the nodes having the given subject. He believed that knowing subject class of more papers can help to classify others more accurately.

Derek Greenee *et al*. [29] proposed a review mechanism for the research themes covered in European Conference on Case-Based Reasoning (CBR) in 2008 and identify the topics that are active at the moment. In their approach, research papers are clustered based on both co-citation links and text similarity. In their experiments on identifying and clustering similar sentences from one or multiple documents of news articles, Marques *et al*. [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. Bader *et al*. [31] proposed a document clustering of scientific texts using citation contexts. In their method, they proved that using citation contexts can provide relevant synonymous and related vocabulary which can help increase the effectiveness of the bag-of-words representation for clustering related scientific texts. For the best of our knowledge there are no proposed methods that cluster research papers in their respective topics using a little information (most important information) for easy retrieval and fast processing.

3. Centroid and Relationship based Clustering Method

In grouping related research papers together, some journals and conferences ask authors to select one or more subjects from a list of subjects when they are submitting the paper. And in some other journals and conferences the task of grouping research paper is done manually and the papers are grouped based on their titles only.

In some journals papers are not arranged in their respective topics but only grouped by volumes and numbers which make it difficult to get related papers. They should be grouped in their topics for easy retrieval. Generally, we believe that a title summarizes the important content of a document [4]. But the effectiveness of this method depends on the quality of the title. In many cases, the titles of research papers do not represent the contents of these documents well. For that we believe that the information of the paper's title, though necessary, is not sufficient to determine a paper's topic. We suggest that clustering academic papers using title and its most related parts can reduce text dimension and produce better further processing. Our algorithm considers paper's title and its related paper information such as its keywords, its related top frequent sentences, and its most similar references in clustering research papers. The proposed approach considers the following features:

3.1. Information extraction

A starting point for computers to analyze unstructured text is to use information extraction [2]. We believe that the most characteristic factors, which describe a research papers, are the meaningful word sequences mentioned within the paper. In order to reduce the dimension of the data for fast processing we have to extract only meaningful important information of the research papers based on their contributions to the research paper. Actually the title of the paper can describe the topic of the paper and Mock [10] believed that terms occurring in the title have higher weights. But in some cases, paper's titles do not describe the topic of the paper well, so in such cases we have to get other information to support the title from the rest of the paper. Other important information of the research paper we are considering in our approach is paper's keyword, top frequent sentences and the most similar references to the paper's title. The details of how this information is extracted are given in Section 4.2.

3.2. Paper relationship

Another feature that is used in this approach is the paper relationship analysis. Mohsen [9] exploited 3 relationships such as citations, common authors, and common references to assign subject to research papers. The research papers can be related to each other and their relations exist in common words in titles, common words in keywords, in common words in frequent sentences and in common referenced papers. So using this information to group similar papers together can reduce the processing time and achieve better results. We consider 4 relationships namely: common words in paper title, common words in paper's keywords, common words in top frequent sentences and common words in most similar references to group similar paper in the same group. A similarity measure is used to measure the similarity between papers, based on their relationships.

3.3. Multi word features

In previous researches, text documents are considered as bag of words [19], word or term has been used in clustering text document but recently methods [16, 18] used multiword features in clustering or classifying textual document.

Clusters produced with multiword features are meaningful than ones produced by single word clustering. A phrase (multi-words) is meaningful to the clustering result only when it is shared by at least a certain number of paper documents. In term of research paper, a phrase is more meaningful than a single word. In this paper, we consider only phrases because they have specific meaning than single word which has a broad meaning. For example the word *information* is a broad word, but phrases such as information *extraction*, *information security*, *information sciences*, *information retrieval*, etc are more specific. We consider only phrases because they are more specific and have specific meaning that single words. Section 4.3 gives more details on this.

3.4. Centroid based clustering

In many previous methods, K-means has been used and 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 research papers. The proposed approach is executed in two modules as shown in Fig. 1. In the centroid selection module, we extract the needed or most important information which includes useful features that will be used in clustering Module. In clustering module we use relationship features in assigning paper to their respective topics.

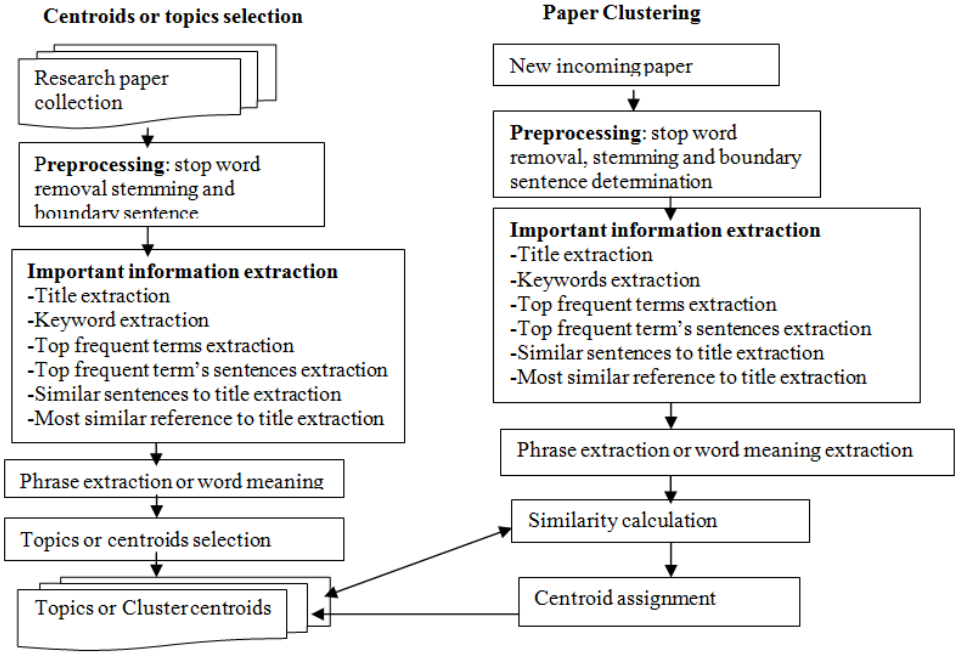

Figure 1. The proposed approach diagram

Each of the process of Figure 1 is explained in the following sections in detail.

4. Feature Extraction and Initial Centroids Selection

In all clustering method, features extraction has to be done so that the best features are selected to be used in clustering process. The proposed approach follows the following steps as mentioned in Figure 1.

4.1 Preprocessing

In this step, we have to prepare the research paper for processing by removing noisy data that can affect the clustering results. Stop words removal, stemming and sentence boundary determination are performed during this stage.

4.2 Important information extraction or Feature selection

We all know that research papers are made of high dimensional text and take long time to process; we want to reduce this high dimension by only considering little information which is very important. In this step we have to select all features that are useful in our clustering method. The paper's title and its keywords are the most significant to the paper and sentences containing title's terms and keywords are relevant so that we can consider them in clustering method.

Generally, the similar sentences to the title contain important terms [8] for paper clustering. Information in which we are interested in bears relationship with the paper's title and keyboards. Such information will frequently appear in a number of sentences and is common among a number of references. In selecting top frequent sentences to be extracted, we first treat each research paper as bag of words to help us to get the top frequent words and then we treat the paper as a sequence of word [16]. The extracted top frequent words will help to get the top frequent phrases or sentences most similar to the paper's title. That means that sentences containing those frequent words are of great concern and using similarity measure only sentences much highly related to the paper's title and keywords are extracted.

When authors write a research paper, they cite some other related papers in which they refer to them which means that their topics are similar or related, we consider this feature as information related to paper's title and use it to support the title's information. In the same way, using a similarity measure, we measure the similarity between paper's title and all references and then only most similar references are considered. The needed information is extracted as follows:

For every paper in the collection, we first extract its title and keywords. We use TF-IDF (Term Frequency-Inverse Document Frequency) to get top high frequent terms or words for each paper. That is

$w_{ij} = Freq_{ij}$ where w_{ij} is the weight of j^{th} term in i^{th} paper document and $Freq_{ij}$ = the number of times j^{th} term occurs in a document. Terms with more weight are selected as frequent words.

Term are selected based on a given threshold or percentage. Then we extract sentences containing those frequent words using algorithm 1 as shown in Figure 2.

Input:
Top frequent words Fw= {Fw_1, Fw_2,……Fw_n)
All sentences of the paper= S_n= {s_1, S_2……s_n)
Output: Selected T sentences

Steps
For all F_w
For all S_n
If Sn_i contains F_w
Extract Sn_i
End For
End for

Figure 2. Top Frequent sentences extraction algorithm 1

Among picked sentences only sentences very similar to the title (at a given threshold) are selected and be used in topics selection.

The similarity between title and each extracted sentences is measured and only those similar to the title are extracted. The title and each extracted sentence are represented as words sets and Dice similarity measure [25] is used to measure the closeness of the sentences to the title.

If T is the title and S is a selected sentence then

$$Sim(T,S) = \arg\max(2(T \cap S)/(T \cup S)$$

The calculated values are normalized into values between 0 and 1 by a maximum value. In the same way, the similarity between paper's title and each reference is computed and only most similar references are selected to be part of cluster centroids.

If T is the title and R is a reference then

$$Sim(T,R) = \arg\max(2(T \cap R)/(T \cup R)$$

Other sentences which are selected are those ones containing keywords. The same algorithm is used to extract sentences containing keywords as shown in Fig. 3.

Keywords K_n = { k_1, k_2…….k_n)
All sentences of the paper= S_n= {s_1, S_2……s_n)
Output:

Selected T sentences

Steps
For all Kn
For all S_n
If Sn_i contains Kn_j
Extract Sn_i
End For
End for

Figure 3. Top frequent sentences extraction algorithm 2

Sentences that contain keywords are all retained and together with extracted frequent sentences similar to the titles (extracted by algorithm 1) are used in initial centroids selection. After this step the concerned important information we have is the title,

keywords, most frequent top sentences and top similar references related to the titles from which we have to select topics or centroids.

4.3 Phrase extraction: Lee *et al*. [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 *sentence* is made. Sentences are the basic unit 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]. 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, we use phrases in our clustering algorithm in order to get meaningful topics or centroids.

We adopted phrase instead of single term because in a text document, a phrase meaning is more meaningful than a simple word meaning. In our method, each selected features such as title, frequent sentences and most similar references in step 2 is viewed as a sequence of words, so that it can be represented as T={w_1 ,w_2 ,w_3 ,...), where w_1 ,w_2 ,w_3 ..are words appearing in T. An ordered sequence of two or more words is called a phrase. Phrase feature is used to determine the topics in which papers are talking about. Paper's keywords are already in the form of phrases as written by authors, we use them as they are, only from titles, references and extracted sentences we have to extract phrases. We adopted Wen [18] method as it was reported to perform well in multiword extraction as follows: Given two sentences S1 and S2 and we want to extract meaningful phrases from them.

Input:
s1, the first sentence
s2, the second sentence
Output: Multiword extracted from s1and s2.

Steps:
s1= {w1,w2,…,wn}, s2= {w1,w2,…,wm'}, k=0
For each word w_i in s1
For each word w_j in s2
While (w_i equal to w_j)
k++
End while
If k>1
Extract the words from wi to wi+k to form a multi-word phrase
k = 0
End if
End for
End for

Figure 4. Phrases extraction algorithm

And then the initial centroids or topics are selected among extracted phrases and keywords as described in the next section. In the next section we will treat both extracted phrases and keywords both as phrases.

4.4 Initial Centroid selection

It is not easy to determine the initial centroids of clusters for text documents. 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 [26]. The buckshot [27] algorithm picks $\sqrt{nk}$ documents randomly from the data set of *n* documents, and clusters them using a clustering algorithm. The *k* centroids resulting from this clustering become the initial centroids. It is known that the clustering algorithms based on this kind of iterative process are computationally efficient but often converge to local minima or maxima of the global criterion function. There is no guarantee that those algorithms will reach a global optimization [6].

We want to have a good set of initial cluster centroids in order to overcome this problem and get better clustering results.

In this step, the frequency of each selected phrase is computed and only top frequent phrases at a certain percentage are selected to represent the cluster centroids.

Usually, short multi-words refer to the general concepts in the documents and long multi-word is a subtopic of these general topics [18]. After extracting multi-words from paper documents, we need to represent the documents using these multi-words. Actually, there are some overlapping among the extracted multi-words where some short multi words are subset of long multi-words, for example if "*remote sensing applications"* and "*remote sensing*" were extracted , we will consider *remote sensing* to be used. Short multi-words are selected to represent cluster.

5. Papers Clustering

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 [25].

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 to perform 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 many previous methods, K-means has been used and has been reported to perform well [6, 13, 24]. 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]. In order to achieve high efficiency of our method, we have also chosen cosine similarity as it has found performing well.

After getting the initial centroids, in the next module we have to group similar papers in the same cluster. As it was done in features selection and initial centroids selection, each research paper has to follow the same steps until it is clustered as shown in Figure 1.

In order to reduce the paper's information, only most important information is considered and only extracted phrases are compared with the selected initial centroids, the paper is

assigned to the clusters is more similar according to the similarity values (at certain percentage).

Considering the content of many research papers, it is very clear that a research paper can belong to more than one topic that is why in our algorithm a research paper can be assigned to more than one cluster.

After selecting K initial centroids each title is assigned to clusters based on a similarity measure between extracted phrases of each paper and the k centroids, then k centroids are recalculated. This step is repeated until all titles and its related information are assigned to clusters. Cosine similarity measure is used to calculate similarity between initial centroids and extracted phrases. Figure 5 on the following page shows the clustering algorithm used.

Input: Research paper Documents N and *K* centroids clusters
Output: Research papers grouped in different K clusters according to their topics.
Steps:
Step1. Select *K* centroids, phrases that have been selected as initial centroids
Step 2. Select the important phrases extracted from research paper *T* from the remaining papers
Step 3. Compute cosine similarities between *T* and *K* centroids
Step 4. Put *T* in the closest clusters (at a certain percentage) and recomputed the centroid.
Step 5. Repeat Steps 2 to 4 until all research paper are processed.

Figure 5. Papers clustering algorithm

The centroids consist of phrases which are central to all titles in the same cluster, thus the closeness in papers' topics are respected. Note that the paper can be clustered in more than one topic and this is done based on the obtained similarity value (at certain threshold).

6. Evaluation of the Proposed Approach

The accuracy of the proposed clustering solution has been evaluated by using the well-known F-measure, Entropy and Purity [14] metrics which have been used in other previous works [6, 14, 21, 28] 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. We run the experiment in interest to know if our reduced paper information can produce better clustering results compared to the ones produced by the full text paper information.

The F-measure is a combination of the precision and recall values used in information extraction. Let n be the total number of paper's titles. If n_i is the number of the members of class i, n_j is the number of the members of cluster j, and n_{ij} is the number of the members of class i in cluster j, then The Precision and Recall for n_i and n_j denoted by P(i,j) and R(i,j) can be defined as follows:

$P(i,j)=n_{ij}/n_j$

$R(i,j)=n_{ij}/n_i$

The corresponding F-measure is computed as

$$F(i,j) = 2((P(i,j) * R(i,j))/(P(i,j) + (R(i,j))$$

The overall F-measure of the whole clustering solution is computed as

$$F = \sum_i \frac{n_i}{n} \max_j (F(\mathrm{i}, \mathrm{j}))$$

Where n is the total number of documents in the data set. In general, the larger the F-measure is, the better the clustering result is [32]. 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.

The calculation of entropy is based on the class distributions in each cluster. This is exactly what is done by Precision metric. The Entropy of a cluster i, denoted E(i) and computed as follows:

$$E(\mathrm{i}) = -\sum_i \mathrm{P}(i, j) \log \mathrm{P}(\mathrm{i}, \mathrm{j})$$

The Entropy of a whole clustering solution, is given by the sum of the individual cluster entropies weighted by the size of the cluster and is computed as follows

$$E = \Sigma_j \frac{|j|}{n} E(i)$$

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

Another metric used is purity [6] .The purity of a cluster represents the fraction of the cluster corresponding to the largest class of documents assigned to that cluster, thus the purity of cluster j denoted by Pu is defined as

$$Pu(j) = \frac{1}{n_j} \max_i (\mathrm{n}_{ij})$$

The purity of the whole clustering result is a weighted sum of the cluster purities is computer as follows:

$$Pu = \sum_j \frac{n_j}{n} Pu(j)$$

We performed our experiments on 200 scientific research papers downloaded from an online open access scientific research library [15]. The library contains many journals and conferences but we randomly selected only 5 journals in the area of computer science and telecommunication such Advances in Computed Tomography (ACT), Applied Mathematics (AM), Advances in Remote Sensing (ARS), Intelligent Control and Automation (ICA) and Communications and Network(CN) to be our test case. Because the downloaded papers were already grouped in their topics, we mixed them and we want our method to group them as they appear in their respective journal topics. Stop word removal and stemming were performed at first for all experiments.

We randomly selected 20%, 40%, 60%, 80% and 100% of our collected papers. We run the experiment 5 times and we recorded the average of F-measure, entropy and purity for each group of interest to know the performance of our method. In the first experiment we treated each research document as bag of words, K-means clustering algorithm and cosine similarity was used. Table 1 shows the results of the experiments.

Table1. Clustering results using FT-IDF method (using bag of words) by considering the whole paper's information

Paper numbers	F-measure	Entropy	Purity
40 papers (20%)	0.828	0.362	0.850
80 papers (40%)	0.817	0.375	0.840
120papers (60%)	0.813	0.419	0.820
160papers (80%)	0.800	0.462	0.810
200 papers (100%)	0.755	0.539	0.780
average	**0.803**	**0.432**	**0.820**

The second test we did it to test our approach using most important information of the research papers as described in our method. In selecting top frequent words, only words having from 1% and above have been selected and for most similar top sentences only sentences appearing 5 times and above have been selected. For the most similar references to the paper's title only those ones having above 0.5 of similarity were selected. At clustering stage, the cosine similarity between selected centroids or topics and extracted phrases of the new papers is computed and the similarity of 0.6 was taken as threshold meaning that the new coming paper can be assigned to more than one topics where the similarity value is 0.6 or above. Table 2 shows the obtained results.

Table2. Clustering results using multi-word and considering only most important information

papers range	F- measure	Entropy	Purity
40 papers (20%)	0.882	0.262	0.900
80 papers (40%)	0.877	0.268	0.890
120 papers (60%)	0.861	0.302	0.880
160 papers (80%)	0.852	0.325	0.860
200 papers (100%)	0.826	0.380	0.840
average	**0.860**	**0.30**	**0.874**

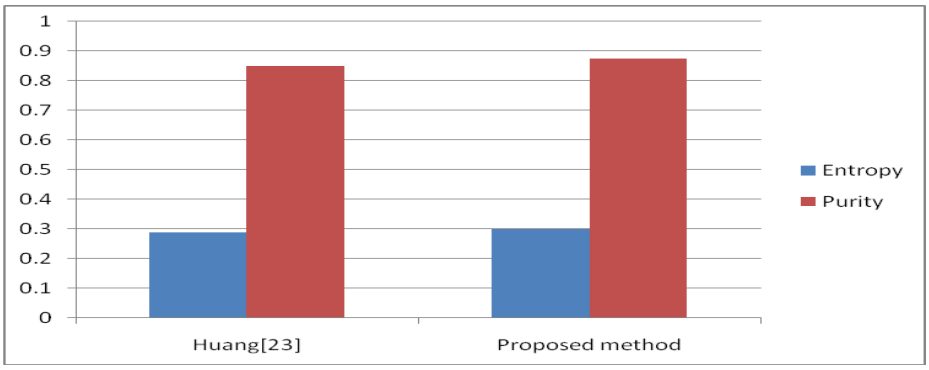

Figure 6. Results comparison with other method

As shown in both tables. Table 1 and 2, it is clear that using most important information, the proposed approach can perform better in most cases. Another

improvement is that it considers most important information with lead to the reduced execution time. The obtained results are also compared with the previous work [23] where we consider only the clustering results on academics papers experiments as it similar to our test case. Figure 6 shows details of the comparison results with previous work.

As it is seen in Figure 6, our method outperforms previous works in term of purity. The best performance of our approach is based on best selection of initial centroids. It achieves 0.19 and 1 of Entropy and Purity respectively for the best case.

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

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 [30].

Our clustering experiment results show an average of 86.0%, 0.30 and 0.874 of F-measure, entropy and purity respectively as shown in Table 2. The clustering results of our proposed approach give the values of 88.2%, 0.26 and 0.9 of F-measure, entropy and purity respectively for the best cases. It outperforms also previous proposed method [23] by 0.024% in terms of Entropy as shown in Figure 6. The best performance of our research papers clustering approach is based on best selection of initial centroids using multi-word feature and by considering common phrases in papers' title, in paper's keywords, in top frequent sentences and in the most similar references while clustering.

7. Conclusion

A centroid and relationship based clustering for research paper is proposed. The method can help reader to get many papers' information in a short time. The new method solves the problem of high dimensional of research paper document by only considering most important information of the paper. Our results indicate that the use of paper's title and keywords when combined with the vocabulary in the full-text of the papers documents is a promising alternative means of capturing critical topics for research papers. The best performance of our clustering approach is based on the best selection of initial centroids using multi word features and considering paper's most important information while clustering. We are planning to extend our approach so that it can be used to group other kind text documents using only their most important information. Key features of our approach are its important information extraction features, paper relationship features, multiword features and its centroid based selection method.

References

[1] S. lee, J.i Song and Y. Kim, "An Empirical Comparison of Four Text Mining Methods", Journal of Computer Information System, vol. 51, no. 1, (**2010**), pp. 1-10.
[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, (**2009**), pp. 60-76.
[3] P. Anick and S. Vaithyanathan, "Exploiting Clustering and Phrases for Context-Based Information Retrieval", Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval, (**1997),** pp. 314-323**.**
[4] B. Endres-Niggemeyer, K. Haseloh, J. Müller, S. Peist, I. Sigel S. A. Sigel, E. Wansorra, J. Wheeler and B. Wollny, "Summarizing information", Simulation of Summarizing, for Macintosh and Windows, Berlin, Springer-Verlag, (**1998),** pp. 307–338.
[5] T. Polajnar, "Survey of Text Mining of Biomedical Corpora", (**2006**), pp. 1-21.

[6] C. Luo, Y. Li and S. M. Chung, "Text document clustering based on neighbors", Journal of. Data & Knowledge. Engineering, vol. 69, (**2009**), pp. 1271–1288.
[7] A. Mustafa, A. Akbar and A. Sultan, "Knowledge Discovery using Text Mining: A Programmable Implementation on Information Extraction and Categorization", International Journal of Multimedia and Ubiquitous Engineering , vol. 4, no. 2, (**2009**), pp. 183-188.
[8] Y. Ko, J. Park and J. Seo, "Improving text categorization using the importance of sentences", Information Processing and Management vol. 40, no. 1, (**2004**), pp. 65–79.
[9] T. Mohsen, "Subject Classification of Research Papers Based on Interrelationships Analysis", In Proceeding of the 2011 workshop on Knowledge discovery, modeling and simulation, (**2011)**, pp. 39-44, New York, USA.
[10] K. J. Mock, "Hybrid hill-climbing and knowledge-based techniques for intelligent news filtering", In Proceedings of the national conference on artificial intelligence, (**1996**).
[11] F. Liu and L. Xiong, "Survey on Text Clustering Algorithm", In Proceeding of. International 2nd conference. Software Engineering and Service Science (ICSESS), IEEE, (**2011)**, pp. 196-199, Beijing, China.
[12] K. Sarkar, "Sentence Clustering-based Summarization of Multiple Text Documents", International Journal. Computer. Science. and Communication. Technology, vol. 2, no. 1, (**2009**), pp. 225-235.
[13] Z. Pei-ying and L. Cun-he, "Automatic text summarization based on sentences clustering and extraction", In proceeding of International. 2nd Conference of Computer Science and Information Technology, ICCSIT, *IEEE*, (**2009**), pp. 167-170, Beijing, China.
[14] E. R. M. Seno and M. d. G. V. Nunes, "Some Experiments on Clustering Sentences of Texts in Portuguese", (**2008**), pp. 133–142.
[15] http://www.scirp.org/journal/CategoryOfJournal.aspx?CategoryID=4, (**2013).**
[16] Y. Li, S. M. Chung and J. D. Holt, "Text document clustering based on frequent word meaning sequences", Data & Knowledge Engineering, vol. 64, (**2008**), pp. 381–404.
[17] K. A. Vidhya and G. Aghila, "Text Mining Process, Techniques and Tools: an Overview", International Journal of Information Technology and Knowledge Management, vol. 2, no. 2, (**2010**), pp. 613-622.
[18] W. Zhang, T. Yoshida and X. Tang, "Text classification based on multi-word with support vector machine", Knowledge-Based Systems, vol. 21, (**2008**), pp. 879–886.
[19] G. Salton and C. S. Yang, "On the specification of term values in automatic indexing", Journal of Documentation, vol. 29, no. 4, (**1973**), pp. 11–21.
[20] K. M. Ganapathiraju, "Relevance of Cluster size in MMR based Summarizer: A Report", (**2002)**.
[21] R. Khoury, "Sentence Clustering Using Parts-of-Speech", International Journal of Information Engineering and Electronic Business, vol. 1, (**2012**), pp. 1-9.
[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, (**2009**), pp. 7764–7772.
[23] A. Huang, "Similarity Measures for Text Document Clustering", NZCRSC, (**2008**), pp. 49- 56.
[24] M. Steinbach, G. Karypis and 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", Journal of Convergence Information Technology, vol. 6, no. 2, (**2011),** pp**.** 22-34.
[26] A. K. Jain and R.C. Dubes, "Algorithms for Clustering Data", Prentice Hall, Englewood Cliffs, (**1988**).
[27] D. R. Cutting, D. R. Karger, J. O. Pedersen and J. W. Tukey, "Scatter/gather: a cluster-based approach to browsing large document collections", in Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval, (**1992**), pp. 318–329.
[28] 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, (**2008**), pp. 641–652.
[29] D. Greene, J. Freyne, B. Smyth and P. Cunningham, "An Analysis of Current Trends in CBR Research Using Multiview Clustering", Association for the Advancement of Artificial Intelligence, (**2010**), pp. 45-62.
[30] L. Jing, K. M. Ng and J. Z. Huang, "Knowledge-based vector space model for text clustering", Knowledge Information Systems, vol. 25, (**2010**), pp. 35–55.
[31] B. Aljaber, N. Stokes, J. Bailey and J. Pei, "Document clustering of scientific texts using citation Contexts", Information Retrieval, vol. 13, no. 2, (**2009**), pp**.** 101-131.
[32] M. Steinbach, G. Karypis and V. Kumar, "A comparison of document clustering techniques", in: KDD Workshop on Text Mining, (**2000**).

Authors

Damien Hanyurwimfura

He received his Masters degree of Engineering in Computer Science and Technology from Hunan University 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.

Liao Bo

He 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.

Dennis Njagi

He received his Bachelor's of Education degree in Mathematics from Egerton University, Kenya in 2000 and a MSc. in Computer Applications Technology from Central South University (CSU), China in 2004. He is currently a lecturer at Jomo Kenyatta University of Agriculture and Technology (JKUAT), department of Information Technology and a doctorate student at CSU. His current research interests include text mining and data fusion.

Jean Paul Dukuzumuremyi

He has got his Master Degree of Engineering in Computer Application Technology from School of Information Science and Engineering at Central South University, China. He is currently a PhD candidate at the same University. His research interests include Digital Image Processing, Computer Graphics, Computer Vision, and their applications.