Text Summarization using Centrality Concept

Ghaleb Al_Gaphari, Ph.D
Computer Faculty Sana'a University, P.O. Box 1247 Sana'a, Yemen

Fadl M. Ba-Alwi, Ph.D
Computer Faculty Sana'a University, P. O. Box 1247 Sana'a, Yemen

Aimen Moharram, Ph.D
Computer Faculty Sana'a, University P.O. Box 1274 Sana'a, Yemen

ABSTRACT:
The amount of textual information available on the web is estimated by terra bytes. Therefore constructing a software program to summarize web pages or electronic documents would be a useful technique. Such technique would speed up of reading, information accessing and decision making process. This paper investigates a graph based centrality algorithm on Arabic text summarization problem (ATS). The graph based algorithm depends on extracting the most important sentences in a documents or a set of documents (cluster). The algorithm starts computing the similarity between two sentences and evaluating the centrality of each sentence in a cluster based on centrality graph. Then the algorithm extracts the most important sentences in the cluster to include them in a summary. The algorithm is implemented and evaluated by human participants and by an automatic metrics. Arabic NEWSWIRE-a corpus is used as a data set in the algorithm evaluation. The result was very promising.

General Terms:
AI Applications, NLP, Text Mining and AI Algorithms

Keywords:
Text Summarization, Text Mining and Centrality Concept

1. INTRODUCTION
Information plays an important role in human daily life in different modern societies. Unfortunately, when large amounts of knowledge are produced and available through the web the process of efficient, effective distribution and accessing this valuable information becomes very critical. In fact, people face disorientation problem because of abundance of such information. Finding specific piece of information in this mass of data requires search engines to perform a remarkable task in providing users with subset of the original amount of information. Anyway, the subset retrieved by the search engines is still substantial in size. For example, at the time of writing the query "Summarization" in Google returned more than 9,090,000 results (as on 30th Jan 2011 from Google). Users still need to manually scan through each single item of the information retrieved by the web search engines until the information of user interest is obtained. This boring task makes automatic text summarization the task of great importance as the users can then just read the summary and get overview of the document .In another word, document retrieval is not sufficient and user need a second level of abstraction to reduce this huge amount of data, user should have text summarization technique. Text summarization is one of the basic techniques in the area of text mining. Text mining is to concern with the task of extracting relevant information, from natural language text, and to search for interesting relationships between the extracted entities [1,23]. In more specific, text summarization is the process of extracting the most important information from a single document / multi-documents and producing a new short version for a particular task and user without losing any important contents or overall meaning from the original document/documents. This process could be seen as a text compression; therefore, text summarization system should define the important parts based on the purpose of the summary or user needs. Text summarization techniques could be classified into two classes based on the way which summarization is going to perform on the input document/documents. Such classes are extractive and abstractive summarizations. The main objective of an extractive text summarization technique is to select the important sentences from the original input text and combine them into a new shorter version. The importance sentences selection process takes place based on linguistic features, mathematical and statistical techniques. The summary generated based on the important sentences from the original input text may not be coherent. But it gives main idea about the content of the input text. While the main idea behind an abstractive text summarization technique is to understand the original input text and then create summaries with its own words. The technique usually, depends on linguistic models to generate new sentences from the original sentences through a process called paraphrasing .The technique includes syntactic and semantic studies for specific language and is useful for meaningful applications. In fact, abstractive text summarization technique is similar to the way a human creates a summary; unfortunately this is still a challenging task for a computer program. As the matter of fact, there are increased demands in developing technologies for automatic Arabic text summarization [14, 15]. Fortunately, there are several research projects to investigate and find out the techniques in automatically summarizing English documents as well as other European languages. Also, there is some software products have been developed for English text summarization such as MEAD summarization toolkit. Unfortunately, there is a limitation in both research papers and software development in terms of automatic Arabic text summarization. The main objective of this paper is to describe results of graph-based centrality algorithm implementation [25]. It is used to capture sentence centrality based on some centrality measures such as degree and lexis ranking. Also, the paper presents a graph representation for clustering documents, where each node of the graph represents a sentence and each edge represents the similarity relation between pairs of sentences. The summarization algorithm is evaluated based on two types of documents that are AFP Arabic newswire corpus provided by LDC, as well as summarization evaluations of Document Understanding Conference (DUC) [24].

2. RELATED WORKS
Over time there have been different methods and techniques to English text summarization and other European languages. Those methods and techniques are associated with single-

document and multi-document summarization. Unfortunately, a few existing projects concerning with Arabic text summarization. The most closely related to this work are surveyed and reported:

A. Haboush et al. 2012 [1] presented and discussed a new model for automatic Arabic text summarization. They stated that the major attribute of their model is the word rooting capability. This attribute enabled the model to be semantic based rather than syntactic based. The meaning behind the root eliminated different derived structures. They reported in their conclusion that they obtained an average of recall (0.787) and precision (0.757) for the resulted summarization.

K. Thakkar and U. Shrawankar 2011 [2] suggested a model that uses text categorization and text summarization for searching a document based on user query. The model uses QDC algorithm for text categorization. The QDC algorithm is evaluated against other clustering algorithms. They stated that by using text summarization after searching the document they save the user's time required for reading the complete document.

P.Vijayapal Reddy et al. 2011 [3] investigated the problem of title word selection in the process of title generation for a given text document by using BMW approach. They stated that they tried to explore the impact of word weigh on Title word selection by using BMW model. They reported that they found F1 measure on Telugu corpus is 1.3 percent less than the F1 measure on English corpus due to Telugu has more complex morphological variations when compared with English.

V. Seretan 2011 [4] presented a novel approach to extractive summarization. The researcher reported that the method produced an abstract for an input document by selecting a subset of the original sentences. The researcher also mentioned that the method based on domain-specific collection. As well as collocation statistics are able to capture the gist of the information content in documents from a given domain, and by the fact that syntactically related co-occurrences represent a better way to model lexical meaning than surface co-occurrence. Finally, the researcher stated that the method has the ability to control the length and detail of the summary produced. Moreover, the work considered, in contrast, only syntactically related word combinations, thus eliminating the need for word sense disambiguation heuristics.

C.F.Greenbacker et al. 2011 [5] introduced an approach to automatic summarization of multimodal documents based on a semantic understanding of text and graphics. They stated that their model enabled them to construct a unified conceptual model that serves as the basis of generating an abstractive summary. They also added that they integrated the knowledge obtained from the graphic with the knowledge obtained from the text at the semantic level. They concluded that their method is able to generate summaries that are more human-like in nature, while not suffering from coherence and other readability issues related to traditional extractive techniques.

N. Nagwani and S. Verma 2011 [6] proposed a summarization algorithm that includes four phases: stop words elimination, frequent term computation, frequent term selection and semantic equivalent terms generation. They reported that all sentences in the document, which are containing the frequent and semantic equivalent terms, are filtered for summarization. They concluded that their experiment result was promising.

H. Yasin et al. 2011 [7] presented an automated Text Summarization System for multiple documents, it based on statistical factors. They stated that, Jacquard's coefficient was used to improve the worth and quality of the summarization. They also mentioned that their experiment was useful and effectual to enhance the quality of multiple documents summarization via Jacquard's coefficient. Finally, they concluded that the system represented steady correlation with the human assessment outcome.

Özsoy et al. 2011 [8] introduced the Latent Semantic Analysis (LSA) method for text summarization. They argued two LSA based summarization algorithm, also, they evaluated both algorithms on two different datasets. They concluded that, both of algorithms perform equally well on both Turkish and English datasets.

N. Zamin and A. Ghani 2011 [9] presented a hybrid approach to Malay text summarization. They indicated that the base system was built based on SUMMARIST and EstSum systems. They also emphasized that using a combination of two techniques enabled the base system to extract the most important sentences from Malay news articles.

H. Saggion 2011 [10] described a language independent multi document centroid-based summarization system. The system was evaluated in the 2011 TAC Multilingual Summarization pilot task where summaries were automatically produced for document clusters in Arabic, English, French and Hindi. The system had a good performance on Arabic and Hindi documents, a medium performance for English, and a poor performance for French.

J. Delort and E. Alfonseca 2011 [11] described the task of update summarization in TAC-2011, which consists of an extension of TOPICSUM. They reported that they have observed that the method performed comparably well for very short summaries in terms of ROUGE-2. Moreover, they executed TOPICSUM on the update set B as a baseline, they shown that it also performed better on shorter summaries.

A. Kogilavani and P. Balasubramani 2010 [12] proposed an approach to cluster multiple documents using clustering method. They produced cluster wise summary based on features profile oriented sentence extraction. They concluded that the generated summary coincides with the human summary for the same dataset of documents.

H. Saggion et al. 2010 [13] presented a series of experiment in content evaluation in text summarization. They reported that they found a weak correlation among different rankings in complex summarization tasks, such as summarization of biographical information and the summarization of the opinions about an entity.

G. Erkan and D. R. Radev 2004 [14] introduced a stochastic graph-based method for computing relative importance of textual units for Natural Language Processing. They evaluated the technique on the problem of text summarization. They stated that the results of applying this method on extractive summarization were quite promising. The main goal of the graph based algorithm is to compute each sentence importance in a cluster and extract the most important sentences to include in the text summary. The process of extraction and combination is based on the concepts of similarity matrix in sentences graph representation [21, 22].

3. PROPOSED ALGORITHM

In this section, we perform the process of computing and combining the sentence centrality scores. Such a process is based on the presence of particular important words and similarity to a central sentence. We also use some measures for centrality such as degree and lexes rank.

3.1.Graph Representation

In a text the sentences are connected to each other. This connectivity can be realized as lexical overlap. In lexical connectivity, two sentences sharing same lexis are connected to each other. This concept is used to compute the sentence importance in a text. Since, a sentence importance in a text is associated with other sentences in the same text. Thus the graph is a suitable technique for representing the relationship and computing the relative importance of sentences by analyzing the graph structure. To implement this concept, the text should be represented as a fully connected graph G= (V, E). Where V is a set of a graph vertices and E is a set of a graph edges. In this study sentences used as the graph vertices at the same time the graph edges represent the lexical similarity between pairs of sentences. As soon as the fully connected graph is constructed the edge reduction algorithms can be used to reduce the graph to include only important edges. The most important edge reduction algorithm is the threshold algorithm. This algorithm eliminates an edge if its weight exceeds some thresholds.

3.1.1. Centrality of a Sentence

A Sentence centrality means the centrality of all words that it includes. The evaluation of a word centrality is to search for the central of the document cluster in a vector space. The centroid of a cluster is a pseudo-document which contains words that have tf×idf scores greater than a predefined threshold [14].

3.1.2. Centroid Based Summarization

The sentences that contain more words from the centroid of the cluster are called central as in Figure 1.

```
input: An array S of n sentences, cosine threshold t
output: An array C of Centroid scores
Hash WordHash;
Array C;
/* compute tf_idf scores for each word */
for i =1 to n do
foreach word w of S[i] do
WordHash{w}{"tfidf"} = WordHash{w}{"tfidf"} + idf{w};
end of each
end of for
/* construct the centroid of the cluster */
by taking the words that are above the threshold*/
foreach word w of WordHash do
if WordHash{w}{"tfidf"} > t then
WordHash{w}{"centroid"} = WordHash{w}{"tfidf"};
end
else
WordHash{w}{"c centroid""} = 0;
end
end for
/* compute the score for each sentence */
for i =1 to n do
C[i] = 0;
foreach word w of S[i] do
C[i] = C[i] + WordHash{w}{"c centroid""};
end
end
return C;
```

Figure 1 Computing Centriod Scores Algorithm

This is a measure of how close the sentence is to the centroid of the cluster [14, 20].

3.1.3. Sentence Salience Concept

A cluster of documents can be seen as a network of sentences which are connected to each other. Some sentences share a lot of information with each other while some others may share a little information with the rest of the sentences. Assume that the sentences which are similar to each other sentence in a cluster are more salient or central to the topic [14].This concept is implemented in this experiment based on computing the similarity between two sentences as well as computing the overall prestigious of a sentence given its similarity to other sentences. The model bag of words is used to represent each sentence as an N-dimensional vector, where N is the number of all possible words in the target language. Cosine similarity measure is used to compute the similarity between two sentences as follows:

$$\cos(x,y) = \frac{\sum_{w \in s1,s2} tf_{w,x} * tf_{w,y} (idf_w)^2}{\sqrt{\sum_{x_i \in x} (tf_{x_i,x} * idf_{x_i})^2} \sqrt{\sum_{y_i \in y} (tf_{y_i,y} * idf_{y_i})^2}} \quad (1)$$

Where $tf_{w,s}$ is the frequency of the word w in the sentence s and idf_w is the inverse document frequency.

A cosine similarity matrix is computed and used for a cluster representation, where each item in the matrix represents the similarity between the corresponding sentences pair. Table 1 shows the similarity matrix which represents a subset of a cluster used in Arab newswire 2004. The same matrix also is represented as a weighted graph where each link represents the cosine similarity between a pair of sentences figure 1.

Table 1: Intra-sentence cosine similarities in a subset of cluster from Arabic Newswire-a (2004)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1.00	0.08	0.04	0.07	0.05	0.03	0.03	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.05	0.02	0.05	0.04
2	0.08	1.00	0.04	0.08	0.16	0.13	0.02	0.00	0.01	0.01	0.00	0.01	0.00	0.06	0.19	0.05	0.04	0.13
3	0.04	0.04	1.00	0.02	0.05	0.03	0.48	0.00	0.05	0.02	0.01	0.00	0.01	0.02	0.12	0.06	0.00	0.01
4	0.07	0.08	0.02	1.00	0.45	0.07	0.00	0.00	0.04	0.00	0.00	0.01	0.00	0.00	0.09	0.05	0.02	0.08
5	0.05	0.16	0.05	0.45	1.00	0.08	0.00	0.06	0.02	0.00	0.03	0.01	0.00	0.16	0.22	0.12	0.03	0.12
6	0.03	0.13	0.03	0.07	0.08	1.00	0.02	0.11	0.00	0.00	0.04	0.00	0.00	0.00	0.02	0.07	0.00	0.00
7	0.03	0.02	0.48	0.00	0.00	0.02	1.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00
8	0.00	0.00	0.00	0.00	0.06	0.11	0.00	1.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.03	0.00	0.00
9	0.00	0.01	0.05	0.04	0.02	0.00	0.00	0.00	1.00	0.01	0.00	0.00	0.00	0.02	0.01	0.09	0.00	0.00
10	0.00	0.01	0.02	0.00	0.00	0.00	0.03	0.00	0.01	1.00	0.00	0.00	0.00	0.02	0.01	0.01	0.00	0.00
11	0.00	0.00	0.01	0.00	0.03	0.04	0.00	0.04	0.00	0.00	1.00	0.06	0.03	0.03	0.01	0.00	0.00	0.01
12	0.01	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.06	1.00	0.00	0.00	0.01	0.00	0.01	0.01
13	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	1.00	0.03	0.02	0.00	0.00	0.02
14	0.00	0.06	0.02	0.00	0.16	0.00	0.00	0.00	0.02	0.02	0.03	0.00	0.03	1.00	0.24	0.02	0.00	0.10
15	0.05	0.19	0.12	0.09	0.22	0.02	0.00	0.00	0.01	0.01	0.01	0.01	0.02	0.24	1.00	0.01	0.01	0.17
16	0.02	0.05	0.06	0.05	0.12	0.07	0.05	0.03	0.09	0.01	0.00	0.00	0.00	0.02	0.01	1.00	0.00	0.00
17	0.05	0.04	0.00	0.02	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	1.00	0.44
18	0.04	0.13	0.01	0.08	0.12	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.02	0.10	0.17	0.00	0.44	1.00

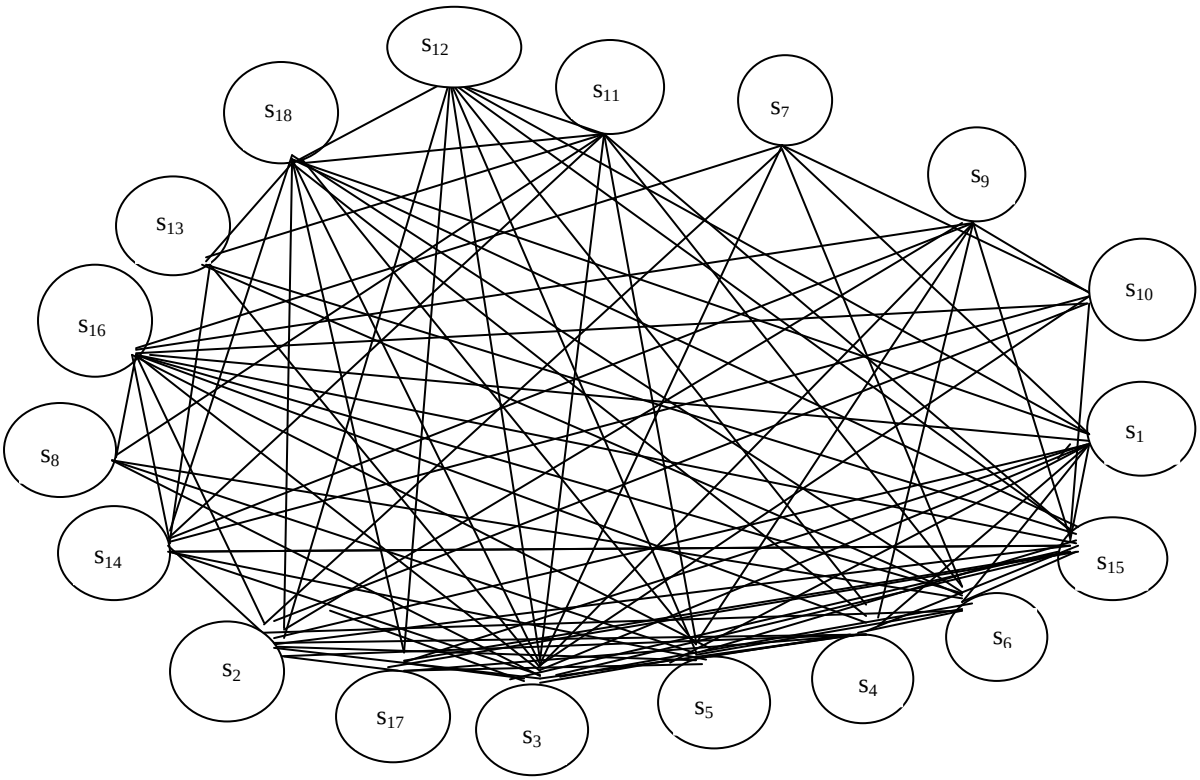

Figure 2: Cosine similarity graph for the cluster in Table 1.

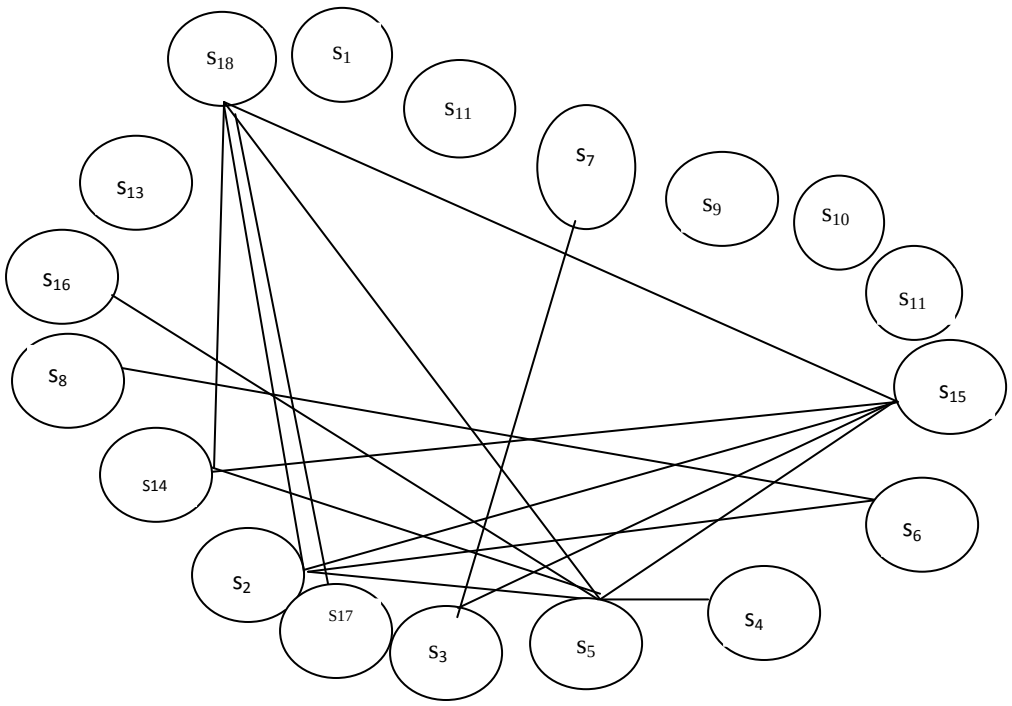

Figure 3: Similarity graphs which correspond to 0.1 for the cluster in Table 1.

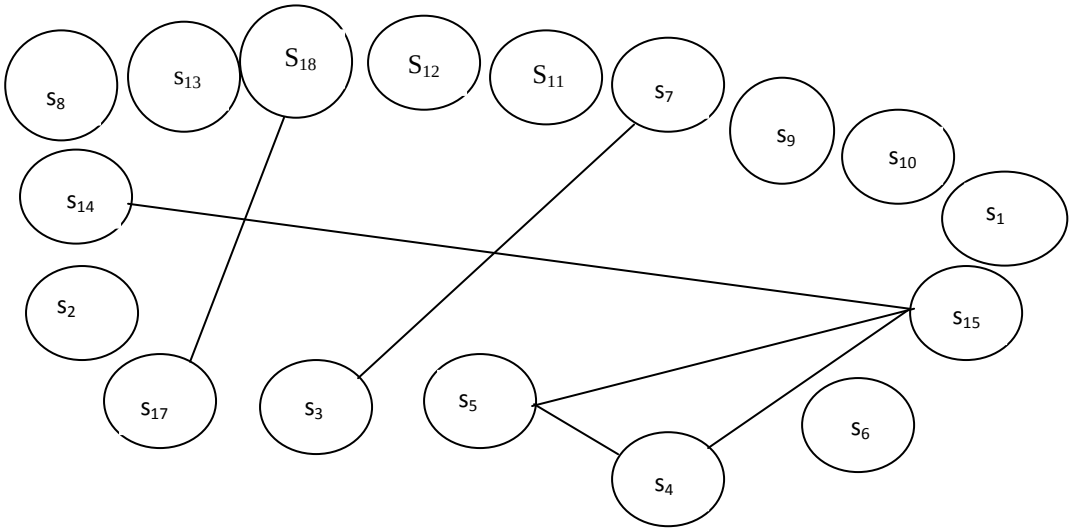

Figure 4: similarity graph which correspond to thresholds 0.2 for the cluster in Table 1.

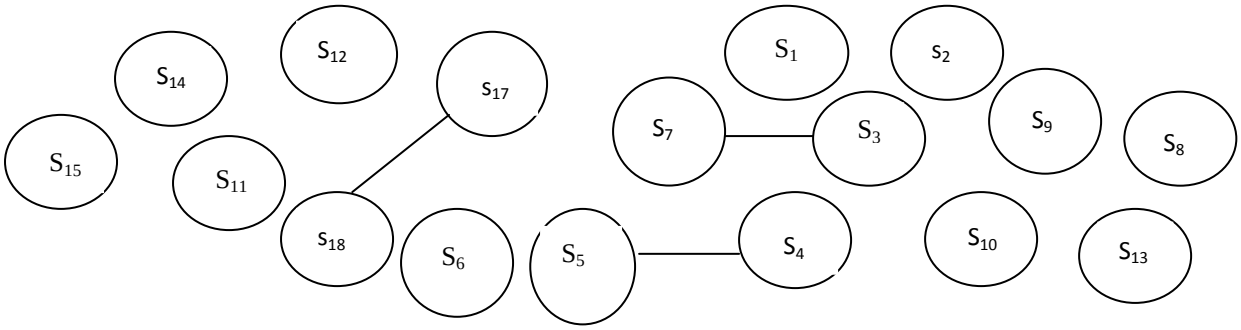

Figure 5: Similarity graph which correspond to thresholds 0.3 for the cluster in Table 1.

3.1.4. Degree Centrality

The degree centrality of a sentence is the degree of the corresponding node in the similarity graph Table1, shows the effect of cosine threshold selection. Too high thresholds may cause losing many of the similarity weights in a set of documents while too low thresholds may cause the weak similarity weights into consideration [18, 19].

Table2: Degree centrality scores for the graphs in Figure 3. Sentence s_{15} is the most central for thresholds 0.1 and 0.2.

Id	Degree(0.1)	Degree(0.2)	Degree(0.3)
S_1	1	1	1
S_2	5	1	1
S_3	3	2	2
S_4	2	2	1
S_5	7	3	2
S_6	3	1	1
S_7	2	2	2
S_8	2	1	1
S_9	1	1	1
S_{10}	1	1	1
S_{11}	1	1	1
S_{12}	1	1	1
S_{13}	1	1	1
S_{14}	4	2	1
S_{15}	6	3	1
S_{16}	2	1	1
S_{17}	2	2	2
S_{18}	6	2	2

4. DATASET AND METRICS:

4.1. Test Collection

The test collection for the proposed algorithm is delivered by the Linguistic Data Consortium (LDC).The LDC provides two Arabic collections, the Arabic GIGAWORD and the Arabic NEWSWIRE-a corpus. The source documents contain meta-data and tags and are represented as UTF–8 files. The dataset contains 100 documents divided into 5 reference sets; each contains 20 related documents discussing the same topic.

4.2. Evaluation

Summary quality and consistency assessment is very difficult, because there is no objective summary. There are two types of summary measures: Form and Content measures. Form measures concern with assessment of text grammar, organization and coherence. Content measures concern with assessment of the percentage of information presented in the machine summary (precision) as well as the percentage of important information omitted from machine summary (recall). Also, there are automatic evaluation measures such as ROUGE. The assessment of this program results was conducted manually and automatically. The manual assessment was based on the text overall responsiveness and the automatic assessment used ROUGE method. For the manual assessment, the human assessors were given the following instructions: Each summary is to be assigned an integer grade from 1 to 5 based on the overall responsiveness of the summary. A text should be assign 5, if it covers the important aspects of the related documents including language fluency and readability. A text should be assign a 1, if it is either insensible, unreadable or contains very limited information from the related documents. The Length Aware Grading Measure (LAGM) was used to normalize the summaries which are out of limit. The (LAGM) is defined as $LAGM = g(1 - \frac{\max(\max(lmin - |s|, |s| - lmax), 0)}{lmin})$ where g is a grade, l_{min} is the lower word limit count, l_{max} is the upper world limit count and |s| is the number of words in the summary. The automatic assessment was based on human created model summary. The summary model produced by the fluent speaker of Arabic language. The RUGE model variations were used[16,17].

5. EXPERIMENT RESULTS

In the resulting summary, all sentences were ranked based on similarity with respect to the centroid. The summary is produced by choosing sentences which are closed to the centroid until the desired bound is reached. A sentence very similar to the centroid appears within the resulting summary before the one is less similar to the centroid. This method gives a coherent summary in terms of processing a single cluster which is centered on specific theme. In this experiment, the acceptable of summary size was between 240 and 250 words.

6. DISCUSSION

By implementing the evaluation measures indicated earlier, the total runs of the program was 5 times .Each run processes 10 documents related to specific theme. The result was shown in Table3, Table4 and table 5 respectively:

Table 3: Human overall and Human LAG responsive scores

Summary Id	Human overall	Human (LAG)
s_1	3.6500	3.6500
s_2	3.7000	3.5458
s_3	4.4500	4.4129
s_4	3.7500	3.7500
s_5	3.9000	3.9000

Table 4: Summary of Precision and Recall for the data set

Summary id	Precision	Recall	F-Measures
s_1	0.8400	0.6287	0.7191
s_2	0.5769	0.5836	0.580
s_3	0.5476	0.5587	0.5531
s_4	0.4889	0.9712	0.9063
s_5	0.7782	0.9146	0.8409

Table 5 : Summary of ROUGE scores for the SBA on the data set.

Summary Id.	Rouge1	Rouge2	Rouge3	Overall
s_1	0.780	0.610	0.461	0.610
s_2	0.670	0.500	0.452	0.541
s_3	0.823	0.653	0.268	0.581
s_4	0.593	0.434	0.346	0.458
s_5	0.549	0.457	0.348	0.451

Table 2 shows the human grading as well as the length aware grading measure (LAG) for 5 different summaries produced by running the programs 5 times on the specified themes. The result in table 1 indicated that the proposed algorithm performed very well. Where the average of the Human grade was 3.89 at the same time; the average of LAG grade was 3.852. The proposed algorithm performed better than the system implemented in [15] as shown in Figure 4.

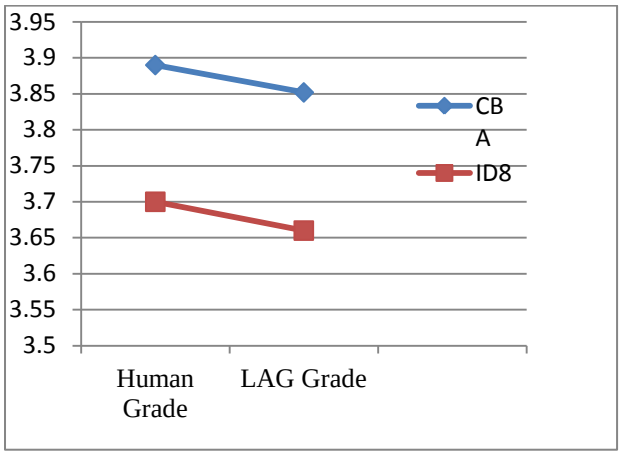

Figure 4: Manual assessment for CBA and ID8

Table 3 illustrates the precision and the recall results. We observed that CBA performed very well. Where the average of the precision was 0.6463 and the average of the recall was 0.7199. The CBA performed better than both the ID8 and the algorithm implemented in [14] as shown in Figure 5.

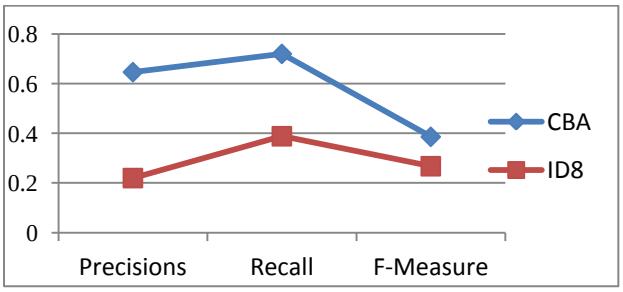

Figure 5: CBA performance along with ID8

Table 5 illustrates the ROUGE results scores for CBA on the data set. Where ROUGE1, ROUGE2, ROUGE3 averages are 0.683, 0.5308 and 0.375. The results show that CBA performs very well. The CBA outperforms the centroid algorithm implemented in [14] as reflected in Figure 6.

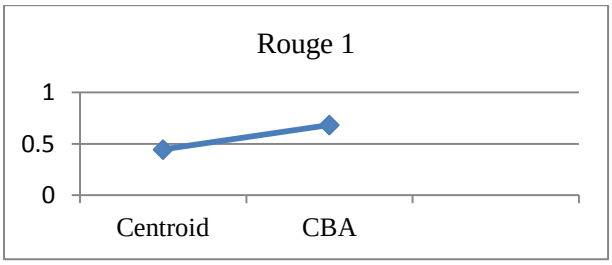

Figure 6: CBA and Centroid performance

7. SUMMARY AND CONCLUSION

In this paper, Centriod-Based Algorithm (CBA) was used for Arabic Text Summarization. A software program was designed, implemented and tested. A real-world dataset was used for testing and validating the software summarizer performance, the result of the experiment was very promising. It shown that the CBA summarizer outperforms other summarizers used for Arabic text summarization performance, so far, such as ID8 and Centriod. On the one hand, the CBA summarizer in this experiment improved the responsiveness scores averages .It raised both the Human (Overall) and Human (LAG) from 3.70 and 3.66 respectively as reported in [15] up to 3.89 and 3.852 respectively as obtained in this task. The CBA raised the F-score from 0.26786 as reported in [15] up to 0.71988 as obtained in this experiment. On the other hand, the CBA raised ROUGE1 from 0.4443 as reported in [14] up to 0.683 as obtained in this experiment.

The future work, the future work will deal with Arabic text summarization using different advanced methods such as ontology, machine learning and other optimization techniques. Neural network, genetic algorithm and association rule. The comparative study among selected algorithms will take place.

8. ACKNOWLEDGEMENT

The author would like to thank anonymous reviewers for their useful comments.

9. REFERENCES:

[1] Haboush, A. ,Al-zoubi, M.,Momani, A. and Tarazi, M. " Arabic Text Summarization Model Using Clustering Techniques". In the World of Computer Science and Information Technology Journal (WCSIT), Vol. 2, No. 3, 62 – 67, 2012.

[2] Thakkar, K. and Shrawankar, U. "Test Model for Text Categorization and Text Summarization ".In the International Journal on Computer Science and Engineering(IJCSE) ,vol.3. No.4 Apr 2011, India.

[3] Vijayapal Reddy, P., Vishnu vardhan,B. and Govardhan,A.," Analysis of BMW Model for Title Word Selection on Indic Script ".In the International Journal of Computer Applications (0975 – 8887) Volume 18– No.8, March 2011.

[4] Violeta, S. ," A Collocation-Driven Approach to Text Summarization". In the TALN 2011 Montpellier, 27 juin – 1erjuillet 2011.

[5] Greenbacker1, C. F. , McCoy1, K.F., Carberry1, S. and McDonald. D., "Semantic Modeling of Multimodal Documents for Abstractive Summarization ". In the Proceedings of the Workshop on Automatic Text Summarization Collocated with Canadian Conference on Artificial Intelligence, 2011, Canada.

[6] Nagwani, N. and Verma,S. ," A Frequent Term and Semantic Similarity based Single Document Text Summarization Algorithm ". In the International Journal of Computer Applications (0975 – 8887) Volume 17– No.2, March 2011

[7] Yasin,H., Yasin,M, Yasin,F.," Automated Multiple Related Documents Summarization via Jaccard's Coefficient ". In the International Journal of Computer Applications (0975 – 8887) ,Volume 13– No.3, January 2011, Pakistan.

[8] Gülçin, Ö., Alpaslan, F. and Çiçekli , İ. ," text summarization using latent semantic analysis ". Master thesis, Middle East Technical University, February 2011

[9] Zamin, N. and Ghani,A. ," Summarizing Malay Text Documents". In the World Applied Sciences Journal12

(Special Issue on Computer Application & knowledge management):39-46, 2011, Malaysia.

[10] Saggion , H., " Using SUMMA for Language Independent Summarization at TAC 2011 ". In the proceeding of the TAC 2011 Workshop November, 2011, National Institute of Standards and Technology Gaithersburg, Maryland USA.

[11] Delort,J. and Alfonseca,E. ," Description of the Google update summarizer at TAC-2011 " . In the proceeding of the TAC 2011 Workshop November, 2011, National Institute of Standards and Technology Gaithersburg, Maryland USA.

[12] Kogilavani, A. and Balasubramani, P. "Clustering and feature specific sentence extraction based summarization of multiple documents ". International journal of computer science & information Technology (IJCSIT) Vol.2, No.4, August 2010.

[13] Saggion, H., Torres-Moreno,J., da Cunha,I., SanJuan,E. and Vel ´ azquez-Morales,P." Multilingual Summarization Evaluation without Human Models " .In the Coling 2010: Poster Volume, pages 1059–1067,Beijing, August 2010.

[14] Erkan, G. and Radev, R., " LexRank : Graph-based Lexical Centrality as Salience in Text Summarization". In the Journal of Artificial Intelligence Research 22 (2004) 457-479.

[15] El-Haj, Mahmoud, Kruschwitz, Chris Fox "University of Essex at the TAC 2011 Multilingual Summarization Pilot".

[16] Haboush,A. , Momani, A. , Al-Zoubi,M. , Tarazi,M. " Arabic Text Summarization Model Using Clustering Techniques", World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 Vol. 2, No. 3, 62 – 67, 2012.

[17] Zamen,N., Ghani ,A. "Summarizing Malat Text Documents ",Wrld applied Science Journal 12 Computer Application and Management,30-46,2011,SSN 1818-4952.

[18] A.Kogilavani1 and Dr.P.Balasubramani2," Clustering and feature specific sentence extraction based Summarization of multiple documents ", International journal of computer science & information Technology (IJCSIT) Vol.2, No.4, August 2010 ,DOI : 10.5121/ijcsit.2010.2409 99.

[19] Perumal ,P. and Nedunchezhian ,R. " Performance Evaluation of Three Model-Based Documents Clustering Algorithms "European Journal of Scientific Research ISSN 1450-216X Vol.52 No.4 (2011), pp.618-628 © Euro Journals Publishing, Inc. 2011.http://www.eurojournals.com/ejsr.htm.

[20] Thakkar K. and Shrawankar U.," Test Model for Text Categorization and Text Summarization ", International Journal on Computer Science and Engineering (IJCSE), ISSN: 0975-3397 Vol. 3 No. 4 Apr 2011.

[21] Yasin, H., Yasin, M. and Yasin, F.," Automated Multiple Related Documents Summarization via Jaccard's Coefficient "International Journal of Computer Applications (0975 – 8887)Volume 13– No.3, January 2011.

[22] Nagwani, N. and Verma,S. "A Frequent Term and Semantic Similarity based Single Document Text Summarization Algorithm ",International Journal of Computer Applications (0975 – 8887) Volume 17– No.2, March 2011.

[23] Haboush,A., Momani,A., Al-Zoubi,M., Tarazi,M. "Arabic Text Summerization Model Using Clustering Techniques ",World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 Vol. 2, No. 3, 62 – 67, 2012 .

[24] Reddy, P., Mahendra,R., vardhan,B., and Govardhan,A. ," Analysis of BMW Model for Title Word Selection on Indic Script ",International Journal of Computer Applications (0975 – 8887) Volume 18– No.8, March 2011.

[25] R. He et al.," Cascaded Regression Analysis Based Temporal Multi-document Summarization ", Informatics 34 (2010) 119–124.