ISSN(Online): 2319 - 8753
ISSN (Print) :2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 1, January 2015

Analytic-Synthetic Processing Of Information as Smart-Based Environment for Text Summarization

Boumedyen Shannaq [1], Richmond Adebiaye[2]

Information Systems Expert& Research Scholar, Ministry of Regional Municipalities and Water Resources, Sultanate of Oman[1]

Program Director & Professor, Computer & Information Systems, College of Business and Technology, Parker University, Dallas Texas, USA[2]

ABSTRACT: The context of this paper is concentrated using the Text Summarization Application (TSA), which is an analysis and design software for complex structures, a smart-based environment implemented with C# programming language. In this paper, important aspects of the implementation of the smart-based environment in TSA are presented to solve problems of highlighting the central keywords of text summarization, thereby presenting a very suitable interpretation of Analytic-Synthetic Processing of Information. It is necessary to utilize the indexing procedure which translates the text in a convenient manner for accurate representation, as well as easy processing of information retrieval. The novelty of this work in using different approaches would be found in the format of extracting different terms as well as in the methods of determining the weightof the terms.

KEYWORDS: Text Summarization, Text Processing, Indexing, Analytic-Synthetic Processing.

I. INTRODUCTION

Modern society places weight on the scope of any professional high standards. It is pertinent to understand that basic knowledge obtained in high school could only be enough for the first 3-5 years of professional career. After this time, the largest part of the information acquired/received could be considered out of date. The essence is based on conception that pure rational knowledge does not need to appeal to experience alone. In order to be "on the crest of a wave," it is imperative to remain competitive with strong ability to constantly update knowledge in all field of activities, i.e. to educate oneself effectively to a large extentsince analytic propositions are misnamed and are logically necessary. The Process of Self-Independent work involves the ability of humans to extract knowledge from existing available information provided in books, and electronic information systems options. The raised problem requires a number of skills and rational work with accurate information. Key aspects include skills in searching information sources; extracting the necessary information from different sources and information processing intended for further use. All of these skills are closely interrelated and require effective methods in order to organize accurate information. Three fundamental concepts are presented in this work: Documentary stream, Collapsing information and Deployment of information [1][2] .

Documentary stream: This is a set of related documents of different types and in various media. Documentary flow is complex and with complex structure. Currently, an individual cannot afford to track the entire volume of documents from any branch of knowledge. Since ancient times, it has been difficult to organize, structure and stream large volume of documentary information. In order to mitigate against uncontrollable volumes, it became necessary to reduce the volume. Thus, began appearance of prototypes of modern methods of collapsing information.

Information Collapsing: This is a method of presenting information in a special economic symbolic form in order to promote its full and rational use.

Deployment of information: Change the physical structure of document meaning as a result of the analytic-synthetic processing, accompanied by a decrease (or increase) in its richness of information.Usually, the features include matching some or all separate word document. In some experiments, it was found to be the most complex representation and very less effective due to inability to have a constant spread and strategically arrangement of words or sentences. In particular, some authors have tried to use a group of words (stylistic, syntax) as features. The aim is to extract the essential keywords in text documents. Essential keywords are the words that represent the main contents of the text, i.e. what the user need to know and remember.

II. RELATED WORK

Many authors and researchers, such as [3] quite convincingly argue that the most likely reason for disappointing results is that indexing techniques based on phrases have the worst statistical characteristics in relation to other methods based on single word, even though their semantic quality could be much higher. [4][5] One of the most common methods of transition to a mathematical model of the document is a "method of using keywords." Keyword is a word or concept in the text, capable in conjunction with other keywords represent text or commands or parameters. Scholars use keywords to reveal internal structure of an author's reasoning. Keywords also reflect the specificity of the paper.
The essence of the mathematical method requires that:

1. For each "classText" creates a list of specific words for that text, then each text can be represented as a vector of frequencies of occurrence of words from this list.
2. There is a problem of "search and selection"of required/essential text words. This is significant because massive amount of information to be processed leads to poorer decision making allowing for urgent selection of keywords which can be distracting.

The aim is to extract the essential keywords in text documents. Essential keywords are the words that represent the main contents of the text, i.e. what the user need to know and remember.

III. INFORMATION EXTRACTION

How does information extraction work? [6] One of the obvious starting point for extracting main information from texts is to read through the texts looking for proper nouns. By using pattern matching with an appropriate lexicon, peoples' names, geographical names, and most importantly, Company names can be quickly identified. Similarly, dates and financial values can be easily identified and picked out.Once this information has been selected of picked out, some structuring is used to help determine the overall meaning of the text based on 'Rule-Based Techniques' [7] [8]. Basically, there are few combinations of keywords "features" used to highlight the important sentences in texts,i.e. those features thatappear within a sentence. In this regard, the sentence would be considered very important. Table 2.1 demonstrates the most common features in Arabic and English languages [9][10][11].

The English Word	*The Arabic Word*	*No*
Its means	يقصد	1
Intended to	المراد به	2
. is	هو	3
Took about	تتحدث	4
Define by specialist	يعرفه أهل الإختصاص	5
Known as	تسمى	6
We can say	يمكن القول	7
In short is	وهي باختصار	8
Describe by expert	يصفه أصحابه	9
Express	يعبر عنه	10
concept	مفهومه	11
Intended to	المقصود به	12
Could be describe	يمكن القول عنه	13
Which means	ويعني بذالك	14
Interpretation/	يفسر	15

DOI: 10.15680/IJIRSET.2015.0401055

Explanation		
Meaning of	معناها	16
Which could means	بما في معناه	17
Narrate/ relate	روى	18
Highlight	يبرز	19
Presented/ Show/ Display	يعرض	20
Contains	إذ يحتوي	21
Diverge	ينقسم	22
The most famous	أشهر	23
appeared	ظهر	24
.Contains	تحتوي	25
Found	وجد	26
Lead to	أدى	27
Shows	توضح	28
Properties	خصائص	29
Factors	عوامل	30
The most	أهم	31
The most Prominent	أبرز	32
Benefits	فوائد	33
Types or Kinds	أنواع	34
Tracking	تتبع	35
So/ Therefore	إذن	36
Represented in	تتمثل في	37
The most important	ومن أهم	38
Case in point is/ For example	ومن أمثلته	39
Refers	يشير	40
Classifies	ويتم تصنيف	41
Includes	يشمل	42
Involves	يضم	43

Table 2.1:Sample of most common features common and used in Arabic and English languages

[12] In addition, we allow and deploy computing algorithm to extract information that is available after punctuation marks, and the information after the punctuation marks produced different meaning as identified below.

1. **( )** or **[ ]** This is for important word
2. **" "** This is for important word or phrase or sentence
3. **.** When sentence is completed/finished
4. **:** It comes after it isspoken/saying or defined
5. **?** The answer comes after question mark
6. **؛**After this, the reasoning becomes materialized or meaningful
7. **!** could beused for great and amazing things/construction of sentences

[13] Other researchers introduce summarization techniques by getting rid the following:

1 - The disjunctive pronouns(ضمائر الفصل) such as: التي , منها , هو

2 - The conjunctions tools(أدوات الربط) such as أيضا , و

3 - Joining styles, if joined to have similar meaning (أساليب العطف إذا كانت المعطوفات متشابهة المعنى).

4 - Computing the long linkages and replacing it to short links. (إحصاء أساليب الربط الطويلة واستبدالها بأساليب ربط قصيرة)

Such as: (نقطة أخرى هامة جدا يجب أن توضع في الاعتبار هي), replace it with (من المهم).

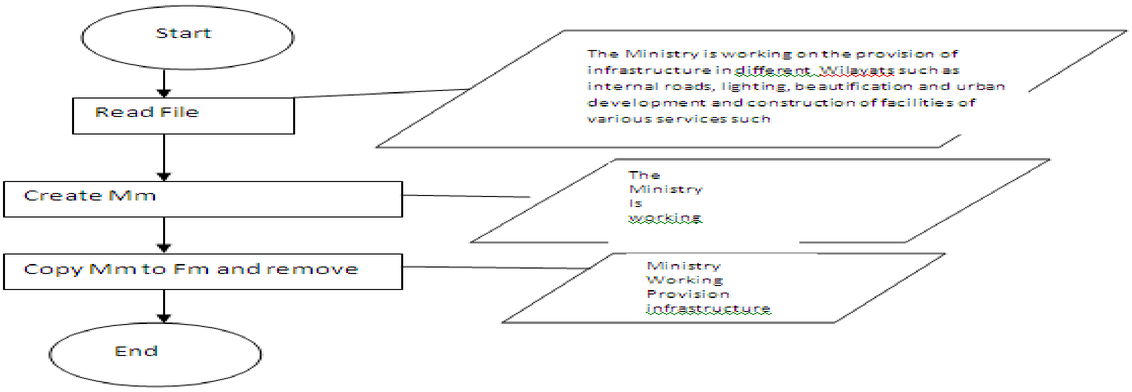
DOI: 10.15680/IJIRSET.2015.0401055

5 - Delete the explained paragraph or between the parentheses.
(حذف الفقرة الشلرحة أو ما بين الأقواس)
6 - Delete the relatingphrase. (حذف جملة الصلة)
7 – Delete speech between the quotation marks.
(حذف ما بين علامات التنصيص وهو الكلام المقتبس لبعض المؤلفين.)
8 - Delete the sentence tandem.(حذف الجمل المترادفة)
Others suggest easy way to summarize how to extract the first line of each paragraph of the text. This may approximates the summary to 25% [14].

In this work we experiment the proposed summarized techniques and found that in some cases, it could show an appropriate result, but it also failed in many areas and few cases.

IV. ALGORITHM

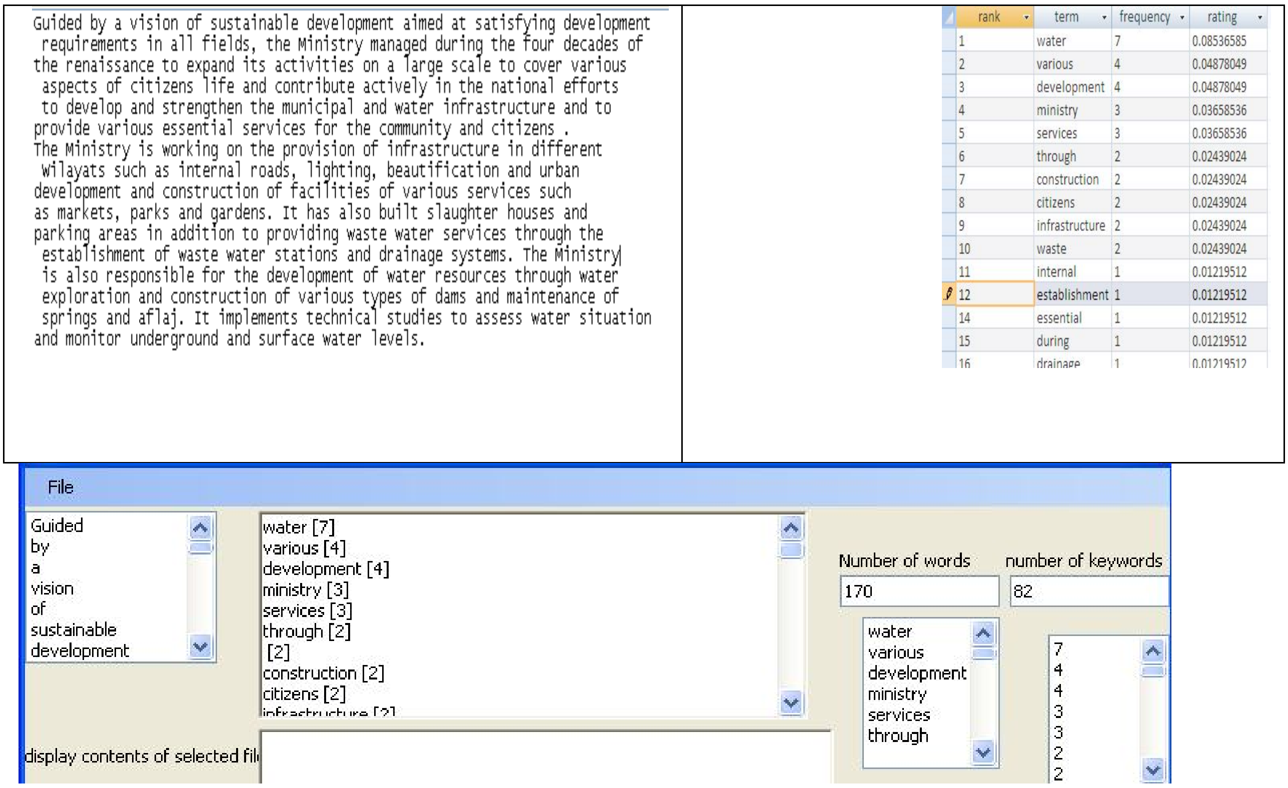
1-Read the file
2- Create main matrix Mm; contains all words including stop words
3-Create featured matrix Fm; contains only keywords (without stop words)
4- Find the central keywords from Fm
5-Find all associated words to the central keywords using Mm.
Figure 3.1demonstrated the implementation of step 2 and 3 of the algorithm


Start
Read File
Create Mm
Copy Mm to Fm and remove
End
The Ministry is working on the provision of infrastructure in different Wilayats such as internal roads, lighting, beautification and urban development and construction of facilities of various services such
The Ministry Is working
Ministry Working Provision infrastructure

I

Figure 3.1: Flowchart information organization

The proposed algorithm is based on Zipf's law of empirical relationship of ranks and frequency [12], which is observed for text having semantic completeness and is analytically represented as:
fr = c,
Where f - frequency of occurrence of words in the text;
r - rank (number) words in the list;
C - an empirical constant.
The rating factor is equal to the relative frequency of occurrence of the word in this text to the total number of significant words in the text (Fm). Thus, the rating of the word in the text – which is the normalized frequency of occurrence in the text.
Example: Figure 3.2, shows a sample text and the process of calculation of Rank, Frequency and Rating as well as a fragment of the developed program used to implement the algorithm.



Guided by a vision of sustainable development aimed at satisfying development requirements in all fields, the Ministry managed during the four decades of the renaissance to expand its activities on a large scale to cover various aspects of citizens life and contribute actively in the national efforts to develop and strengthen the municipal and water infrastructure and to provide various essential services for the community and citizens .
The Ministry is working on the provision of infrastructure in different Wilayats such as internal roads, lighting, beautification and urban development and construction of facilities of various services such as markets, parks and gardens. It has also built slaughter houses and parking areas in addition to providing waste water services through the establishment of waste water stations and drainage systems. The Ministry is also responsible for the development of water resources through water exploration and construction of various types of dams and maintenance of springs and aflaj. It implements technical studies to assess water situation and monitor underground and surface water levels.

rank	term	frequency	rating
1	water	7	0.08536585
2	various	4	0.04878049
3	development	4	0.04878049
4	ministry	3	0.03658536
5	services	3	0.03658536
6	through	2	0.02439024
7	construction	2	0.02439024
8	citizens	2	0.02439024
9	infrastructure	2	0.02439024
10	waste	2	0.02439024
11	internal	1	0.01219512
12	establishment	1	0.01219512
14	essential	1	0.01219512
15	during	1	0.01219512
16	drainage	1	0.01219512

```
Guided
by
a
vision
of
sustainable
development
```

```
water [7]
various [4]
development [4]
ministry [3]
services [3]
through [2]
 [2]
construction [2]
citizens [2]
infrastructure [2]
```

Number of words: 170
number of keywords: 82

```
water
various
development
ministry
services
through
```

```
7
4
4
3
3
2
2
```

Figure 3.2: Fragment Of The Developed Program

Rating = frequency / number of keywords
The rating of water equal to; 7/82 =0.08536585
The “water “keywords in this example represents the central keyword of the text, therefore the summarization process is performed based on “water”.
The visualized structured in figure 3.3demonstrates this implementation:

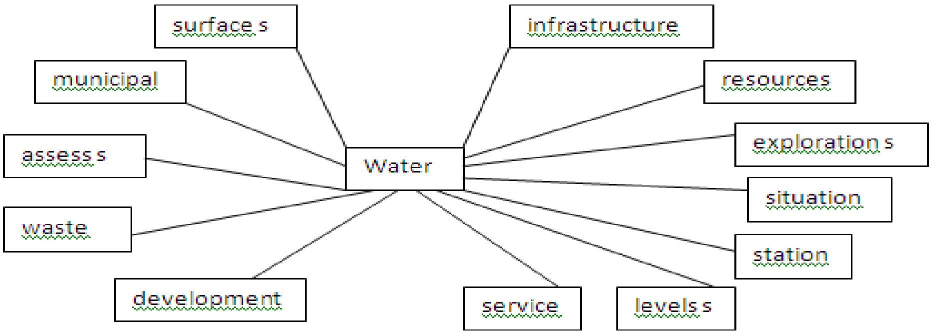


Figure3.3: “Visualized structure of “Water”

The final result of summarization process of text demonstrated in figure 3.2:

strengthen the municipal and water infrastructure, to providing waste water services through the establishment of waste water stations, the development of water resources through water exploration to assess water situation through surface water levels

The final results show that 'water' is one of the main problems raised by the ministry. The application is implemented using C# programming Language which is also capable of summarizing English and Arabic texts. For a word list, sorted in descending order, according to the words in the texts, dates from the rankings (as well as the frequency of occurrence) are hyperbole for all natural-language texts, regardless of their scope and content. Interests are only the first few percentages of the entire list of words, as they constitute its conceptual core. As a result of the summarization process, we obtain "Water" texts examples of high-level generalizations of large fragments of texts, which can serve as headings and subheadings in the documents, the phrases in the annotations, and abstracts for papers and so on.

V. CONCLUSION

The processing of documents is provided with new qualities that contribute to its identifications, retrieval and document distributions. Through summarization, a new document is obtained and outlined. In an ideal case, the problem of summarization of documents create series of arbitrary texts, smaller in volume than the original text, while still maintaining its core. The basic types of information summarization, such as indexing, bibliographic description annotations, reviewsand overview and analysis activities, including many other techniques describe instantly any text, based on the frequency analysis of occurrences of words in the text. However, this is clearly insufficient to assess the document in the collection. Model TF * IDF allow mathematical, vector model text highlights list of keywords. The proposed algorithm and its implementation as well as the experiment demonstrated the possibility of applying the model to real-world examples. This forms accurate heuristic techniques to improve the selection of keywords and expanding the list of 'stop-words'. The construction of the vector is not considered a consistent order of words but the context represents an important semantic component of the text. From the perspective of possible improvements of the method, we note: Automation of parameter selection by discarding non-semantic load bearing words offered by rating factor; considering the location of the construction of words in a document, we recommend merge, split texts. This is important for a possible construction necessary for quality vector representation.

REFERENCES

[1 Boumedyen , Hines J., Adebiaye R., "BWA_Thesaurus as knowledge management strategy for Arabic and English Tourism Webpage's " WCAS the international Conference ICKMARS-2012.

[2] Salton, G. and McGill, M.J.," Introduction to modern information retrieval", McGraw-Hill, 1983.

[3] Lewis, D.D., An evaluation of phrasal and clustered representations on a text categorization task. In Proceedings of SIGIR-92, 15th ACM International Conference on Researchand Development in Information Retrieval (Kobenhavn, DK, 1992), pp. 37–50., 1992

[4] T. Joachims ," A probabilistic analysis of the rocchio algorithm with TFIDF for text categorization", In Proc. of the ICML'97, 143–151, 1997.

[5] Boumedyen ,Victor ," Clustering the Arabic Documents (CAD)", Universal Journal of Applied computer Science and Technology (UNIASCIT), Vol 1 (1), 2011, 05-08 .

[6] Dagan, I., Karov, Y., Roth, D., Mistake-driven learning in text categorization. In Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (Providence, US, 1997), pp. 55–63., 1997

[7] Antoine Dahdah, a glossary of grammar, ninth edition, Beirut. Lebanon Library Publishers, Page 4

[8] d. Nabil Abu Haltam - Nazmi sentences, Nabil Zein - Zuhdi Abu Khalil. Encyclopedia of Arabic language. Jordan - Amman g Dar Osama for Publishing and Distribution 2003 Page 8.

[9] Zine full Alkhwska, rules of grammar - Alexandria. First Edition, fulfillment house for a minimum printing and publishing, 2002 page 10.

[10] Muhammad Ali Kholi. Methods of teaching Arabic, edition 2000, Amman - Jordan. Dar Al-Falah for publication and distribution in 1997. Page 68

[11]Boumedyen Shannaq , "Adapt Clustering Methods for Arabic Documents", American Journal of Information Systems, 2013, Vol. 1, No. 1, 26-30 Available online at http://pubs.sciepub.com/ajis/1/1/4 © Science and Education Publishing DOI:10.12691/ajis-1-1-4.

[12]Boumedyen Shannaq ," Investigating the Distribution of Arabic and English Keywords and Their Progress Over Different Text File Formats", American Journal of Computing Research Repository, 2013, Vol. 1, No. 1, 1-5 Available online at http://pubs.sciepub.com/ajcrr/1/1/1 © Science and Education Publishing DOI:10.12691/ajcrr-1-1-1.

[13] Boumedyen Shannaq," Methods and Algorithms for Searching Arabic Name Entity", International Journal of Computer Applications, Volume 82 - Number 8, 2013.

[14] Boumedyen Shannaq ,"Using Russian and English ontology in expanding the Arabic query", Universal Journal of Applied computer Science and Technology (UNIASCIT), Vol 1 (1), 2011, 05-08