Science Arena Publications
Specialty Journal of Electronic and Computer Sciences
ISSN: 2412-7485
Available online at www.sciarena.com
2018, Vol, 4 (4):7-15

# Clustering-Based Text Improvement and Summarization Based On Collective Intelligence Algorithm

**Hossein Baloochian[1], Maryam Nazari[2*]**

[1] Computer - Software Group, University of Safir Danesh, Ilam Branch, Iran.
[2] Master Student, Computer - Information Technology Group, University of Safir Danesh, Ilam Branch, Iran.

***Corresponding Author**

***Abstract:*** *This research aims to improve and summarize the text based on clustering based on collective intelligence algorithm. The algorithm that is calculated in this way is based on the binary particle aggregation algorithm. Each particle size in this algorithm is measured with a fitness function, but instead of using the speed equation, the new particle position is calculated. What has been done in this study is to provide a general hybrid model of the two TD -IDF algorithms along with the PSO multi-factor clustering Which covers the main body. The proposed method is based on the method of weighting the TF-IDF mechanism, Mr. Salton, which uses the repetition of the document's words and queries to calculate the weight. The main idea is to specify a coefficient for each semantic transference and refer to the two terms that are involved in this transfer. Then, in counting the frequency of the words of these sentences, the coefficient is multiplied by the frequency. The net PSO algorithm provides optimal clustering solutions. To increase the speed and precision of the system, we use two local searches based on the composition particle structure and, at the end, we see several percent improvement over the previous work.*

***Keywords:*** *Text Summarization, Clustering, TF-IDF Algorithm, PSO.*

## INTRODUCTION

Considering the increased volume of text documentation and web-based resources development to meet users' information needs, other information retrieval techniques are no more efficient. Although studing a large amount of literature is very difficult for users while searching and having a brief summary of the topic can be very useful (Khosheghbal, 2014), text compression process whose result may prove useful to the user is called summarization. Text summarization results in more and higher-speed use of resources and a consequent availability of more useful information for the user. When it comes to automatic text summarization, the differences between computer-generated and human-generated text summaries and the potential efficiency of automatic summary compared to human-made summary are what pop up to our minds. Humans are naturally able to understand the concepts embedded in the text and the relationship between them, while it is extremely difficult to do it by machines. On the other hand, humans have a different view towards the same text summary according to their information context knowledge level (Dorigo and Stutzle, 2004). In the present study, we first need to find the link between the nominal terms of the existing corpus[1]. As we know,

[1]text corpus

there are sources in English that contribute to deriving the parts of speech[2]. In this paper, it was tried to pre-process the corpus with stop terms and the remaining text will contain nouns. The SPC can be used to analyze the simultaneous occurrence of the terms and compare it with a threshold value. If it would be bigger than this value, it will be defined as a double-name term. Here we begin extracting vectors. To create context term vectors, we consider a text window with a size of CW=20. Now for each name existed in the corpus, we will consider the simultaneous occurrence of the other terms and this term. Then, we store those terms that occurred the most in its neighborhood; therefore, we will have a vector for each of these nouns. After extracting these vectors, we can compare the similarity of the two terms in the vector space, and if the two vectors in the vector space are close together, it indicates that the two terms are interrelated and used in a similar text. It should be noted as well that the two terms may appear once in the text and the SPC value become 1, however, it is insignificant. Therefore, we need to consider the weight associated with each term among which the weight of TF-IDF is the most important, which is the frequency of repeating one term in ISF[3]. Due to the increased importance of electronic text resources in the world wide web, the need for an accurate, rapid, and easy access to information becomes more apparent. Moreover, the considerable human force that can be employed for other tasks, may be directed to facilitate access to information. In order to properly and quickly find the user's intended documents, reading a full huge text is inappropriate. The automatic summarization system bridges this gap; its advantages are as follows: 1) The summarized text size is manageable, i.e. the machine can provide the summary with respect to the level of the user's desired compaction, 2) Its contents are predictable; i.e., it can be determined that each part of the summary is related to which portion or portions of the main text. The purpose of this study is to achieve a balanced system compared to previous methods to better summarize the text on the web and employ collective intelligence algorithms to summarize the text based on the clustering technique. It is assumed that the proposed method will improve the text quality and summarization speed. The application of velocity equation instead of computing the new position of the particle, has a positive effect on improving the proposed method.

**Previous Works**

In the source (Moshki & Analouyi, 2009), the new clustering-based method was proposed for multi-document summarization of Persian texts. In this method, after preprocessing the text including determining terms and sentences boundaries, text unification, common terms elimination, and multiple text elements identification, the main summarizing process starts. In the summarization step, the sentences are clustered first and then the sentence that are most relevant to other sentences is selected for each individual cluster. In the final summarization step, the sentences are inserted in the final summary according to the chronological order of the texts (news). The results show that in most cases, the output of the proposed summarization system produces an acceptable summary (over 80 %). In the source (Khosheghbal, 2014), summarization concepts and its variants as well as summarization steps and summary extraction methods are expressed. Next, we will discuss and evaluate all types of summary. Finally, the effective factors in summarization and its areas along with the benefits of this process are presented. In the source (Radev et al., 2004), a multi-document summarization system, i.e., MEAD is presented. In this system, for each sentence in a cluster of relevant documents, the three features, e.g., the center, location, and similarity with the title are computed and a linear combination of these features is employed to identify the important sentences. In their method presented in 2012, Gupta et al. combined a single-document summaries using sentence clustering techniques to create multi-document summaries. The process starts with the creation of a single-document summary using a scoring method. The sentences are then clustered using syntactic and semantic similarity between sentences to determine parts of the text to be included in the summary. Finally, the summary is generated by extracting a sentence from each cluster. In the source (Ferreira et al., 2014), a summarization system is

[2] Word functions or roles

[3]Inverse Sentence Frequency

introduced. In this method, the summarized sentences are determined by graph-based clustering algorithm that employs statistical and linguistic similarities. The algorithm transforms the text into a graph and then identifies the original graph sentences using Text Rank. The sentences are then grouped based on their similarities.

**Description of the proposed method**

In the proposed algorithm, each particle has a binary value; its value is measured by a fitness function; but instead of using the velocity equation, the new particle position is calculated as follows.

$$p_{IJ}(t) >= \frac{1}{1+\exp(-v_{i,j}(t))} \rightarrow X_{ij}(t+1) = 0$$

$$Else\ X_{ij}(t+1) = 1$$

Additionally, this study is based on the collection of Latin resources for the first chapters for which the concepts and methods used will be described. Then, for the analysis chapters, the methods of other scientists who have implemented similar works to this study will be utilized to which new approaches will certainly be added. For the experimental evaluation of the presented system, the generated summary will be evaluated on the DUC2002 data set. The data set contains 567 documents in the form of 59 clusters. The documents are news reports from newspapers and news agencies such as the Financial Times, Associated Press, and Wall Street Journal. For each collection of documents, two ideal summaries will be produced by experts in the form of one cluster. This application may serve other purposes too. In text filtering, especially emails, and generally any application that is intended to organize documents or a specific selective and comparative distribution of documentation, these automated question-answering systems may prove useful. The text classification is a natural language processing problem and can be viewed as an attribution of the unclassified documents to one or more pre-defined categories based on their content. Automatic classification operates at a much better time and expense. Moreover, different methods and algorithms have been applied for classifying the text, which function differently in terms of precision and calculation. Feature selection is a suitable procedure for reducing the problem size as well as the computational cost. In text classification, the features should be chosen so that they can be useful in distinguishing a text from its shadow. The selected features for a class should not overlap selected features for other classes. However, there are many procedures for feature selection, such as the use of information gain, document test and frequency; however, these procedures are not efficient enough for solving the multi-class problems. Since many of these features have too much redundancy, the need for a procedure to reduce the dimensionality would be optimal. The method proposed to solve this problem employs the UNI and ICF parameters associated with each term and then chooses a term that has an ICF smaller than the specific threshold limit and a UNI bigger larger the specified threshold limit. By scoring word repetition in fewer classes and a higher number of its occurrences in a class, these parameters attempt to select those words that are more representative of a specific class as an exponential feature.

**Simulation parameters**

One of the most important issues in indexing which identifies the role of words in terms of their effect as keywords of the text is the term weight. In this stage, using different weighting patterns, a weight is attributed to each extracted term or the phrase. This weight indicates the effect of the term on the original content of the text compared to other terms used in the text. Based on the primary hypothesis, it was attempted to identify the two authors who published their works in two different fields. In this study, we generally examined the two authors who have published ten articles in the fields of literature and astronomy. In the first step, based on the first section of the articles of the two authors (preface and keywords), we extracted ten specialized keywords, with complete repetition in each article and other articles for each author, and set up a library. Following the simulation process, we will attempt to use a combination of twenty terms

extracted from the works of the two authors as a feature extraction and weighting key to determine the author.

Text extraction algorithm:

- Start
- Text recall
- Application TF-IDF algorithm
- Application of Text classification algorithm
- Application of PSO
- Conclusion

In text classification, like any other classification problems, it is necessary to determine the appropriate training set. Term weighting in text classification plays an important role in enhancing the classifier performance. TF-IDF approach is a popular term weighting and feature selection method in text classification. In this study, we used data for checking and weighting based on twenty key terms. Moreover, the TF-IDF weighting criterion was used to construct the document space vector. To apply the semantic features of the terms and to add to their connotation (meaning) in the weighting, their impact should be somehow determined by applying them into this criterion. Here, we will provide two proposed changes in both the local and global dimensions of the TF-IDF criterion, both obtained experimentally. The main idea in this section is to specify a coefficient for each semantic transfer which will be applied to the binomial involved in this transfer. Then, in calculating the term frequency of these sentences, this coefficient should be multiplied by the frequency. Similarly, to pay more attention to the importance of a term in transferring concepts at the text level, which is one of the major weaknesses of criteria such as TF-IDF, the grammatical distribution of the term at the text level is also reflected in term frequency count in a text. For a text in the experiences tested in this study, we have run the following steps to count the term frequency: step 1: In this test, we have assigned a frequency coefficient to each semantic transfer in a text according to Table 1. This coefficient is determined based on the intuitive guesses offered for semantic transfer as well as the average occurrence of each transfer in the entire document set which will be describe later. These coefficients have been obtained based on several replications of the data recovery test with the proposed algorithm in improving the Measure-F test precision criterion.

**Table 1**: The empirical coefficients obtained for semantic transfers based on their average occurrence in the data set.

| Transition | Coefficient | The average occurrence in the Medline set |
|---|---|---|
| Continue | 2.3 | 0.135 |
| Smooth-shift | 1.4 | 0.17 |
| Retain | 0.6 | 0.28 |
| Rough-shift | 0.51 | 0.415 |

One of the most important intuitions in this test is the lack of precision in this kind of coefficient assignment to disorganized texts in terms of distribution of concepts at the text level. For example, in a text with a high number of Continue semantic transfer occurrences, many text terms enjoy a high frequency coefficient. Thus, this text and especially its terms, ultimately will have a high TF-IDF value. In answer to many user queries, this text is considered a good candidate. Whereas, disorganization and inappropriate distribution at the text level, in terms of concept dispersion at its level, is an important cause for the inappropriateness of such a candidate. This claim has been proven in terms of tests and observation of different texts. Even in summarized or indicative texts that a concept has spread in a small volume of sentences throughout the text, the number of semantic transfer occurrences is normal in which a high number of semantic transfers will never occur. Of course, the occurrence of semantic transfers like Retain and Rough-shift appears to be self-evident; this is mainly due to the futility of a large number of sentences in a long text which do not play a role

in transferring the meaning and implications of that text. Step two: in this test, we first count the total number of semantic transfers and the total number of semantic transfers at the level of a text. Then, based on the repetition rate of a transfer in the text and the total number of text transfers, we will normalize it based on eq. (3 - 1). The computation of this normalized frequency is derived from the sentence-term matrix:

$$F \times TrF = \text{calculated weight for each semantic transfer}$$

$$TrF = \sum \backslash li = 1 \text{ here l is the number of text sentences for each specific transition}$$

$$ITF = \quad Log \frac{\text{Number of text sentences}}{\text{trf}}$$

If a transfer is repeated, its value is reduced to a logarithmic scale. In fact, this equation is the main TF-IDF criteria equation. The weight allocated to each transfer in the text is calculated by multiplying the criteria obtained from the above equation by the value obtained experimentally through repeating Measure-F test on the obtained Medline document set. The value of this auxiliary parameter is calculated based on the Table 2 for each transfer that is experimentally achieved.

**Table 2**: Empirical coefficients used for semantic transfer from multiple test on the data set.

| Transition | Coefficient |
|---|---|
| Continue | 1.8 |
| Smooth-shift | 1.3 |
| Retain | 0.8 |
| Rough-shift | 0.7 |

Furthermore, the PSO algorithm creates close-optimal solutions for clustering. To increase system velocity (speed) and precision, we use two local queries based on the hybrid particle structure. That is to say, a K-Means step is used to improve cluster centers. Since in some cases, zero may be given to some centers. And we drop this cluster center and randomly assign a new cluster center instead. A local search is also done to estimate the feature weight. That is to say, we perform a Gaussian mutation with the probability of 0.15 on each feature, and the weight change algorithm will remain in case of increasing the fitness function.

## Simulation Findings

The proposed method will result in improving the quality and summarization speed of the text. The same feature vector terms used in the learning algorithm are employed to construct the classifiers. In articles, various methods have been used for weighting these features, such as binary TF-IDF weighting, etc. In binary-weighting, 1 means the existence of a term (feature) in the training text, while zero means non-existence of a term in the training text. In TF weighting, only the number of raw frequency of the terms is considered. In TF-IDF weighting, for a word such as $W$ in a text such as $d$, the aggregation of IDF(w) and TF (w, d) are calculated. Given the use of this algorithm and the background used in the same cases, we will observe the exact term weighting in classification and term weighting. The quality of the classification is based on evaluation criteria feedback. Assigning weight to the files is done according to the class where they fall. In this algorithm, in addition to assigning weight to the features, the weight of files is calculated in all classes. The file will be transferred to the class in which it has more weight. Instead of calculating the new position of the particle, the velocity equation has a positive effect on the improvement of the proposed method. In this study, we have use a hybrid particle structure to combine two suitable weight estimation functions for the features and to determine the cluster centers at the same time. The system consists of four artificial data sets and three genuine data sets.

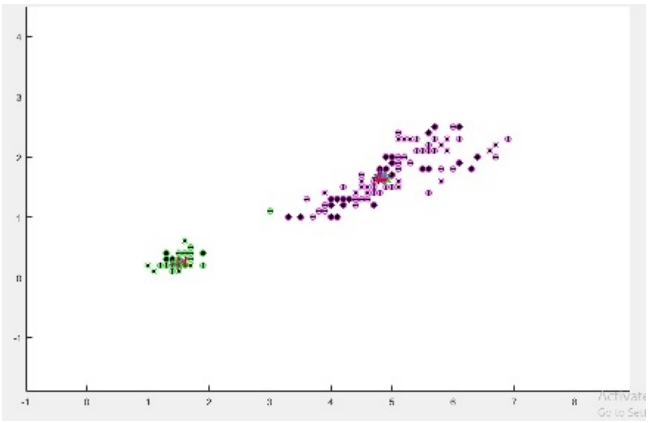

**Figure 1.** genuine data sets for 150 actual data in two clusters.

The results are obtained from the performance of the systems on the test data set that results from the average precision of ten times implementation. As we can see, the proposed method could nicely perform the clustering operation on the data set and has very few faults (with a low error rate).

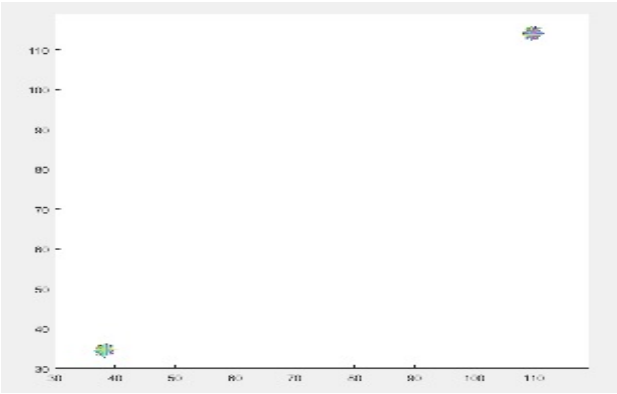

**Figure 2.** Genuine data sets for 110 actual data in two clusters.

In this thesis, a new method has been presented for term weighting in language processing applications, information retrieval and etc. This new method is based on two semantic and linguistic perspectives. One of the most important reasons for considering these views in term weighting, is the dependence of the text defining features, namely the terms (words), the intrinsic and implicit features, such as the meaning of the term as well as its grammatical role. As we have seen, many of the current term weighting methods rely only on statistical properties of words in texts, i.e., their repetition frequency.

This dependence alone causes an inadequate performance in the initial word clustering, which are clearly seen in the results obtained from different approaches. To fill this gap, noting the inherent language features in this thesis, by selecting the TF-IDF method and changing the TDF local weighting parameter from the semantic and linguistic point of view, it has been attempted to improve the effectiveness of this method. The main reason for choosing a method such as TF-IDF as the basis criterion, is its simplicity of implementation and calculation. These two are the main reasons for the popularity of this method among weighting criteria. In this path, two main ideas were investigated to simultaneously cover meaning and language grammar. To cover the meaning and add text semantic features into term weighting, a linguistic theory, i.e., centrality theory was investigated and used. One of the most important and proven properties of centrality theory so far is its ability to distinguish local consistency[4] and prominence[5] in a text. The main claim of this theory is to

[4] Also, cohesion, integrity

distinguish semantic transfers between sentences of a text. Thus, in the aforementioned thesis, the focus was on changing the TF parameter, which is a local weight parameter. On this basis, two steps were followed. The first step was based on the average repetition of each semantic transfer in a series of texts being processed, a weight was assigned to each semantic transfer, and then this weight was multiplied by the weight of the terms in which a particular semantic transfer occurred. The results of this step were not favorable enough. Therefore, the calculation of each semantic transfer weight was taken into account in the next sub-step. In the second sub-step, the weight of each semantic transfer in a text was calculated based on the given formula. In this weight, the coefficient (factor) 1 is multiplied for each text set which is calculated only once. Then for each word, the value 1, i.e., seen only once somewhere in the text, is multiplied by the semantic transfer weight of the sentence, covering that word.

## Conclusion

In this study, based on the authenticity of the mission hypothesis, a combination of algorithms was used. The following is a detailed comparison with which the main criterion is precision. As we have seen, the proposed method has solved two categories of challenges to a large extent: the challenges facing the statistical methods, e.g., disregarding meaning and language features, low precision, and low efficiency are three challenges facing methods with an ideational approach to term weighting. Since the proposed method does not depend on any initial knowledge sets, no additional processing is needed due to the use of learning techniques or dictionaries. It is loyal to the initial data set and does not expand or alter it.

**Table 3:** Comparison of the algorithms

| Data type | K-Means method | FAPSOACO method | The proposed method |
|---|---|---|---|
| Artificial | 61 | 68 | 75 |
| Genuine | 65 | 70 | 77 |

According to the comparison in the above table between previous algorithm and the new proposed algorithm, we have seen improvements compared to the previous algorithm. Natural language processing (NLP) is one of the major branches in the extensive field of artificial intelligence as well as linguistics knowledge. The major efforts done in this field, was to automate the understanding of concepts expressed in a natural language. To put it more clearly, natural language processing is to use computers to process spoken and written language. We can use it to translate languages, use web pages and written information databases to answer questions, or to negotiate with devices such as for consultation. Language post-processing was followed by new movements, however, it had never been able to solve the problem. Natural language processing is a subcategory of artificial intelligence, which deals with the development and application of computational models for language processing. In this field, there are two main disciplines: 1) Perception, which deals with processes that extract information from the language (such as natural language perception, information retrieval), and 2) Production, which deals with processes using language to transmit information. Generally speaking, in natural language processing, human and computer are closely related to each other. The computer is mentally embedded in the human brain. Consequently, a vacuum system is formed for which the human being is the major organizer. In fact, the purpose of natural language processing is that the computer would be able to analyze and evaluate human language and even be able to produce a natural language. The main purpose of natural language processing is to present language computational theories using algorithms and data structures in computer science. Written and spoken applications are the most important areas of natural language processing application, which allows for finding specific information in a text or translation of a text to another. In recent years, there has been growing research that realization of these goals requires

[5] Also, salience

extensive knowledge of language. Therefore, in addition to computer science scholars, the knowledge of the linguist will help to access services and different language processing systems, such as customer communication systems via telephone or student training systems. Text processing includes four levels: lexical processing, morphological processing, syntactic processing, and semantic processing. Each of the many applications of text processing, including information retrieval, summarization, perception, production, question and answer, knowledge extraction from texts, etc. is performed at one or more levels according to the range and complexity. What was done in this study was to present a general combined model, which constitutes two TD-IDF algorithms and a PSO multi-agent clustering. Finally, a few percent improvement has been observed in comparison to previous works, which represents the integrity of the research. There are a number of future works on the continuation of the activity on semantic and grammatical weighting which will be mentioned a little later to attract the attention of proponent. One of the challenges facing the proposed method, is the centrality method dependence on document set reference language. The weighting criteria such as TF-IDF are by no means dependant on the reference language in terms of absolute dependence on statistical properties such as frequency. Nevertheless, language theories such as centrality theory or even intuitive guesses that are presented on the preference of grammatical functions can change according to the reference language. The proposed rules for the theory of centrality have been especially investigated in this paper on English language. The centrality theory was evaluated in several other languages which were mentioned in the literature review section. Notwithstanding, the assessment of it has not been offered for e.g., Persian language so far. To use the proposed method for a language such as Persian, theoretical or experimental evaluation of centrality theory on the Persian language needs to be performed. Then, the extension of the proposed method and converting it into a multi-language approach is an open future work. One of the open future works, is the extension of centrality theory capability in certain applications such as term weighting to the entire document set. As it was mention earlier, the only ability of centrality theory in recognizing local consistency and prominence is proven. The proposed intuitions are expressed on the priority of semantic transfers with respect to each other based on the same proofs. Nonetheless, as for the ability of centrality theory to distinguish global consistency and prominence in a set of documents, good results have not been achieved. However, attempts have not been made by using available tools in computer science and math to prove this hypothesis, i.e., whether the centrality theory has the capability to recognize global consistency and prominence. One of the leading tasks is to use a machine learning method to distinguish a number of semantic transfers occurring on a continuum, as a chain, shows global consistency and prominence. Of course, this learning should take place in a specific application, like multi-document summarization for local consistency and specifying the local preference of semantic transfers from its usage in single-document summarization. Therefore, the priority of semantic transfers or semantic transfer chains can be determined in a set of documents. Now for assigning weight to semantic transfers, one step further leads to a global-weighting of semantic transfers or a chain of semantic transfers. In this step, the use of machine learning methods is very effective. In one of the pioneering works, we can parameterize the coefficients used in this thesis. As we have seen, for assigning weight to semantic transfers and grammatical functions, some coefficients were used based on the required document set being processed. It seems that the use of soft computing methods, such as the fuzzy method in the calculation and deployment of these coefficients can help to adapt the proposed method to the document set and largely improve the effectiveness of the method.

## References

1. Ferreira, R, de Souza Cabral L, Freitas, F, Lins, R. D, de França Silva G, Simske, S. J. and Favaro, L. (2014), “A multi-document summarization system based on statistics and linguistic treatment,” Expert Systems with Applications, vol. 41(13), pp 5780-5787.

2. Gupta, V. K. and Siddiqui, T. J. (2012), "Multi-document summarization using sentence clustering," Fourth international conference on intelligent human computer interaction
3. Khosheghbal, S., (2014). "Automated Text Summarization", The First National Conference on Electronic Technology Advances in Electrical, Electronic and Computer Engineering, Electronics, Khayam Electric University.
4. Kylie Bryantm,Genetic Algorithm and the traveling salesman problem, Hervey mudd college, 2000.
5. Moshki, M. & Analouyi, M. (2009)."Multi-document summarization of Persian texts using a cluster-based method", the first national conference on software engineering in Iran, Rudehen, Islamic Azad University, Rudehen Branch
6. Radev, D., Jing, H., Stys, M. and Tam, D.(2004), "Centroid-based summarization of multiple documents," Information Processing and Management, vol. 40(6), pp 919-938.