Jaccard Coefficients based Clustering of XML Web Messages for Network Traffic Aggregation

Dhiah Al-Shammary
College of Computer Science and Information Technology
University of Al-Qadisiyah
Iraq.
d.alshammary@qu.edu.iq

Recived : 1\4\2019 **Revised : 18\4\2019** **Accepted : 22\4\2019**

Available online : 6 /5/2019

Abstract

This paper provides static efficient clustering model based simple Jaccard coefficients that supports XML messages aggregator in order to potentially reduce network traffic. The proposed model works by grouping only highly similar messages with the aim to provide messages with high redundancy for web aggregators. Web messages aggregation has become a significant solution to overcome network bottlenecks and congestions by efficiently reducing network volume by aggregating messages together removing their redundant information. The proposed model performance is compared to both K-Means and Principle Component Analysis (PCA) combined with K-Means. Jaccard based clustering model has shown potential performance as it only consumes around %32 and %25 processing time in comparison with K-Means and PCA combined with K-Means respectively. Quality measure (Aggregator Compression Ratio) has overcome both benchmark models.

Keywords: Jaccard Coefficients, XML, Clustering, Similarity Measurements

1. Introduction

Web communication protocol is a significantly main medium to transfer information over the internet represented in XML messages [4, 8]. Recently, different kind of XML based protocols have been developed adding potential services to the network like video conference and surveillance in addition to secured and private peer to peer communication [14,1, 5]. However, these protocols still require potential improvements and contributions to solve having the situation of network bottlenecks and congestions.

1.1. Problem and Motivation

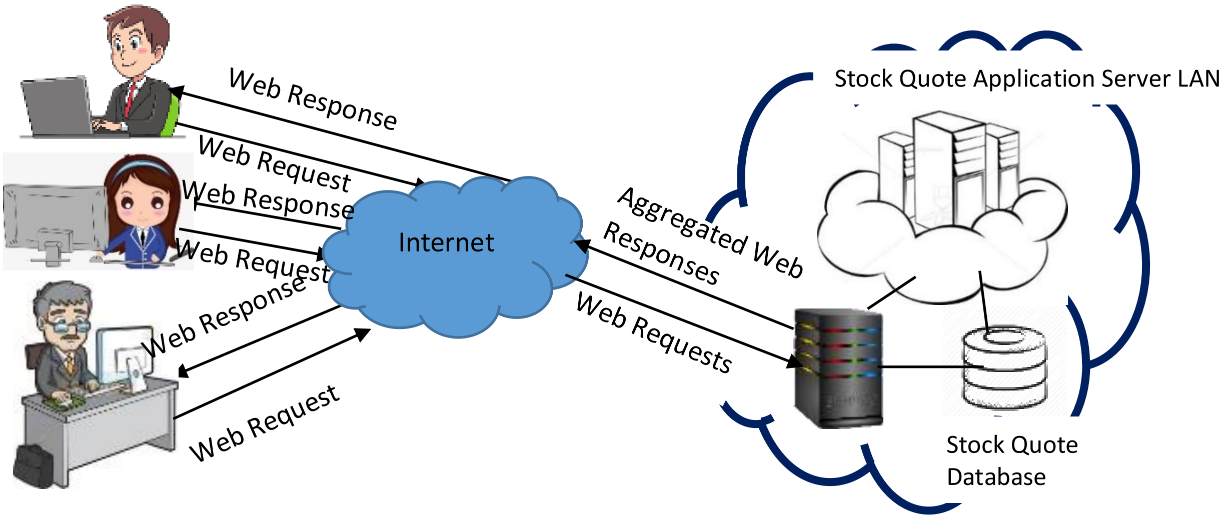


Figure 1: Jaccard Clustering Support for Stock Quote Web Service

XML based Web messages representation has degraded the network performance by their highly redundant information creating high network traffic [1, 5, 7, 16]. As a result of XML significant network volume, users have suffered from usual network bottlenecks and congestions [7, 2]. In many cases, network response time is very slow delaying user's services [5, 7, 8, 9]. This fact has motivated researchers and industrial engineers to adopt potential techniques that trying to reduce network traffic and speed up services response time. Web messages aggregator is one of the most recent proposed contributions attempting to catch group wise messages redundant information and creating one compact structure for a number of messages [10, 4]. However, potential aggregation depends on the similarity level of grouped messages. Therefore, a similarity measurement model is required to guarantee high aggregation results. An efficient clustering would support this requirement and improve Web services to endpoints over the network.

1.2. Contribution

This paper proposes an efficient clustering model for XML messages based on Jaccard coefficients similarity measurement. This similarity measurement has very low complexity that can guarantee a high network response time. Technically, Jaccard coefficients is calculated for two messages only. However, the proposed model has utilized this similarity metric to find the similarity of more than two messages and group them together. As a result, Web aggregator do not need to worry about finding similar messages and would only aggregate highly similar messages that can produce efficient network traffic reduction. Figure 1 illustrates a suggested scenario for supporting Web aggregators over the cloud by storing Stock Quote endpoints big data in a compact aggregated version over the cloud saving huge storage space.

1.3. Evaluation Strategy

An efficient evaluation strategy is designed and implemented by testing the proposed model efficiency. This is achieved by delivering its grouped messages results to the XML based aggregator developed by [2] and check its impact on network traffic reductions (Compression Ratios) and Processing time. The same approach is taken with other clustering models such as K-Means and PCA combined with K-Means as they have been selected as a benchmark for our proposed clustering technique. Technically, Jaccard clustering model has shown tremendous results in comparison with other models as it consumes potentially less processing power and at the same time overcome them by reducing even more network traffic.

1.4. Organization of the Paper

The rest of this paper is organized into five sections. Section 2 illustrates the related work and how other researchers attempted to cluster Web messages. Section 3 explains in details the proposed Jaccard clustering model. Section 4 presents a deep analysis and evaluation results of the proposed model. Finally, section 5 summarizes a conclusion of overall the paper.

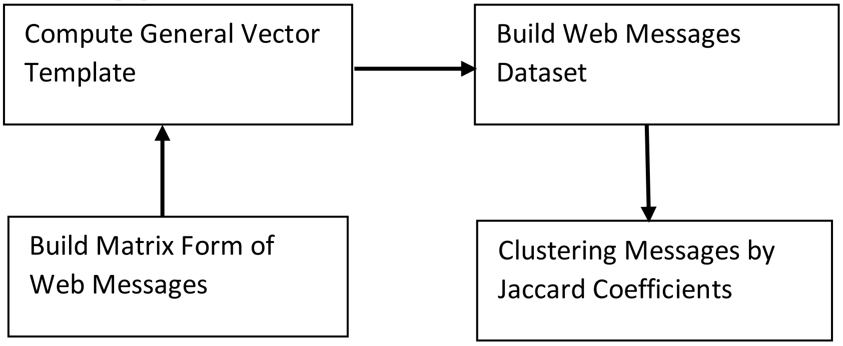
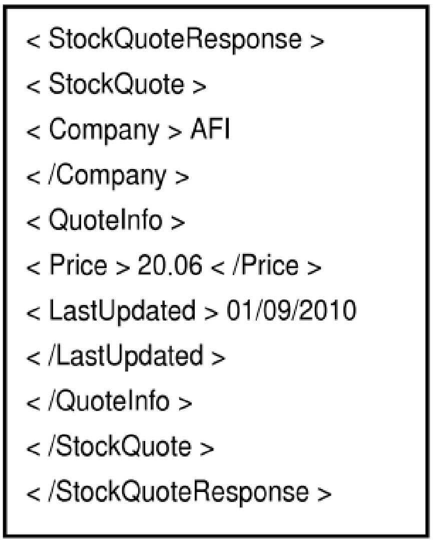

Compute General Vector Template
Build Web Messages Dataset
Build Matrix Form of Web Messages
Clustering Messages by Jaccard Coefficients

Figure 2: Main Jaccard Clustering Model Components

2. Related Work

Several studies have been achieved to provide potential clustering that can group numbers, messages or documents together based on how close they are to each other. A main direction of the developed clustering models is dedicated to cluster XML messages and documents. At the same time, another main direction of clustering techniques is dedicated to work on numbers and cannot work on text unless these textual documents are transformed to a related binary form.

K-Means is one of the best known model based on Euclidian distance function. K-Means selects messages to be centroids and then search the closest messages to be linked with centroid message by having the shortest Euclidian distance. Another study is achieved to enhance K-Means by introducing Principle Component Analysis as a pre-processing to K-Means. PCA is efficient in summarizing data bringing the main information into beginning of its results placement. Data with two-dimensional numbers, PCA will transform them into the same size two-dimensional numbers with %80 of the main information summarized in the first two columns. This fact has been utilized to summarize data and cluster them using only the first %80 information located in the first and second columns. This has created more complexity in comparison with K-Means with better clustering results.



```
< StockQuoteResponse >
< StockQuote >
< Company > AFI
< /Company >
< QuoteInfo >
< Price > 20.06 < /Price >
< LastUpdated > 01/09/2010
< /LastUpdated >
< /QuoteInfo >
< /StockQuote >
< /StockQuoteResponse >
```

Figure 3: Example of Web Textual Message

XML messages have gained high interest of researchers and industrial engineers to cluster them directly working on the XML structure and format. Yongming [8] has developed a novel XML based clustering model. They have utilized vector space model with some extensions to capture the similarity parts of XML messages. Their proposed model works on a transformed XML dataset created by extracting features like recording leaf nodes path and weight. The empirical results have shown enhancements to the clustering job. Another model is presented by Hwang and Gu [10] by catching large XML structures and find out their frequencies inside the XML tree and other related tress. Their model works on other traditional features like XML nodes path. They cluster XML messages based on the higher similar frequencies of large XML structures in addition to tags and items. Their results have proven that they enhanced the clustering requirement and provide better results.

3. Proposed Model

The proposed technique consists of two main concepts: preparing XML messages into binary representation based on a generic numeric template and clustering messages based on Jaccard similarities using a predefined cluster size. Figure 2 shows the main processing components.

3.1.Dataset Transform

Datasets usually have a specific format that is required for a specific clustering model. In this research, the first step is building XML trees of the considered XML messages. Figures 3 and 4 represent a real XML message example and its XML tree representation respectively. Then, XML matrix form is generated by level order traversing XML messages. Figure 5 shows the XML matrix form of XML tree. XML time series representation would be calculated using their XML matrix forms.

Distinctive items of all XML messages are collected to build their generic vector template in preparation to build the binary dataset of XML messages. Equation 1 illustrates the general shape of the generic vector template.

$$T = \{item_1, item_2, item_3, \dots, item_n\} \qquad (1)$$

Equation 2 illustrates the overall representation of dataset based on the weighted XML items (frequencies).

$$X_1 = \{w_1, w_2, w_3, \dots, w_n\}$$

$$X_2 = \{w_1, w_2, w_3, \dots, w_n\}$$

$$X_3 = \{w_1, w_2, w_3, \dots, w_n\}$$

$$\vdots$$

$$X_m = \{w_1, w_2, w_3, \dots, w_n\} \qquad (2)$$

The Term Frequency with Inverse Document Frequency (TF-IDF) is applied to calculate wi as illustrated in Eq. 3.

$$w_i(d) = tf_i \times log\frac{D}{df_i} \qquad (3)$$

Where

- tf_i is the frequency of XML item inside the document d.
- df_i is the frequency of documents having instances of XML item.
- D is the total number of involved documents in the whole dataset

3.2.Jaccard Coefficients Grouping

Jaccard Coefficients also known as Jaccard Index is a statistical factor that is used to find out similarity of two sets. It is widely applied to work on two sets only (Web messages). In this paper, Jaccard coefficient is applied as a base similarity function to work on a group of messages. The proposed clustering strategy is applied by selecting one of the involved XML messages as a centroid for each targeted cluster. Then, based on a pre-defined cluster size, the proposed model allocates messages based on their Jaccard similarity with the centroid message. Equation 4 illustrates the Jaccard Coefficient similarity metric of two sets A and B.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \qquad (4)$$

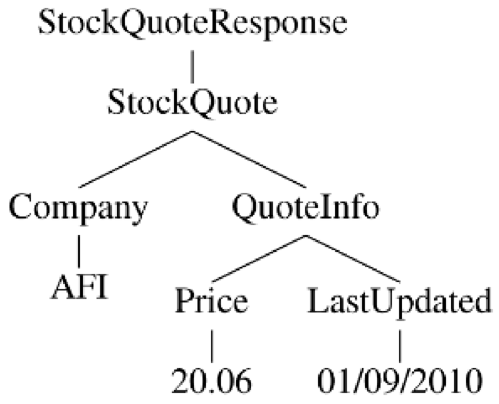
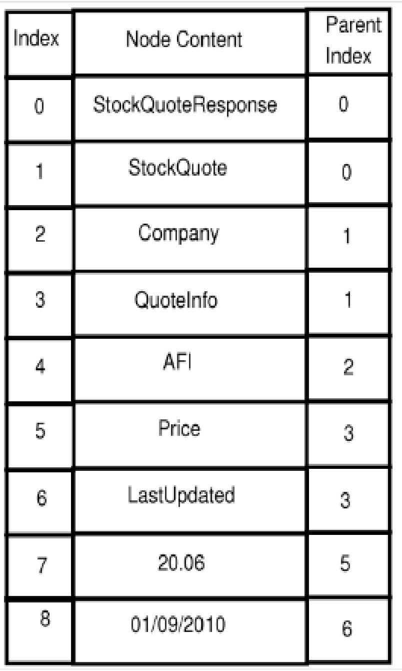


Figure 4: XML Web Message Tree

XML messages first transformed to generate dataset numerical representation. Equation (4) can be simplified into Eq. 5.

$$J(A,B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \qquad (5)$$



Index	Node Content	Parent Index
0	StockQuoteResponse	0
1	StockQuote	0
2	Company	1
3	QuoteInfo	1
4	AFI	2
5	Price	3
6	LastUpdated	3
7	20.06	5
8	01/09/2010	6

Figure 5: Matrix Form of XML Web Message

The proposed model clusters messages based on their generated numerical representation. First, randomly selected messages are distributed into clusters as centres. Then Jaccard coefficients are calculated to the rest of messages in relation to centre points. Finally, messages are allocated to clusters based on the most similar Jaccard metric against the centre message.

4. Experiments and Discussions

The experiments to testify the proposed Jaccard coefficients clustering has used XML dataset published by [2] that consist of four groups

Algorithm 1: Jaccard Clustering Distribution

```
Input: Vector numerical representations of Web messages

Output: Jaccard based metric Groups of Web messages

    1. Select messages randomly with the same number of required clusters
    2. Distribute selected messages as centre points of clusters
    3. For each not clustered message do
        a. Calculate Jaccard Coefficient of current message with all other centre points
        b. Allocate current message to the best match centre point cluster
```

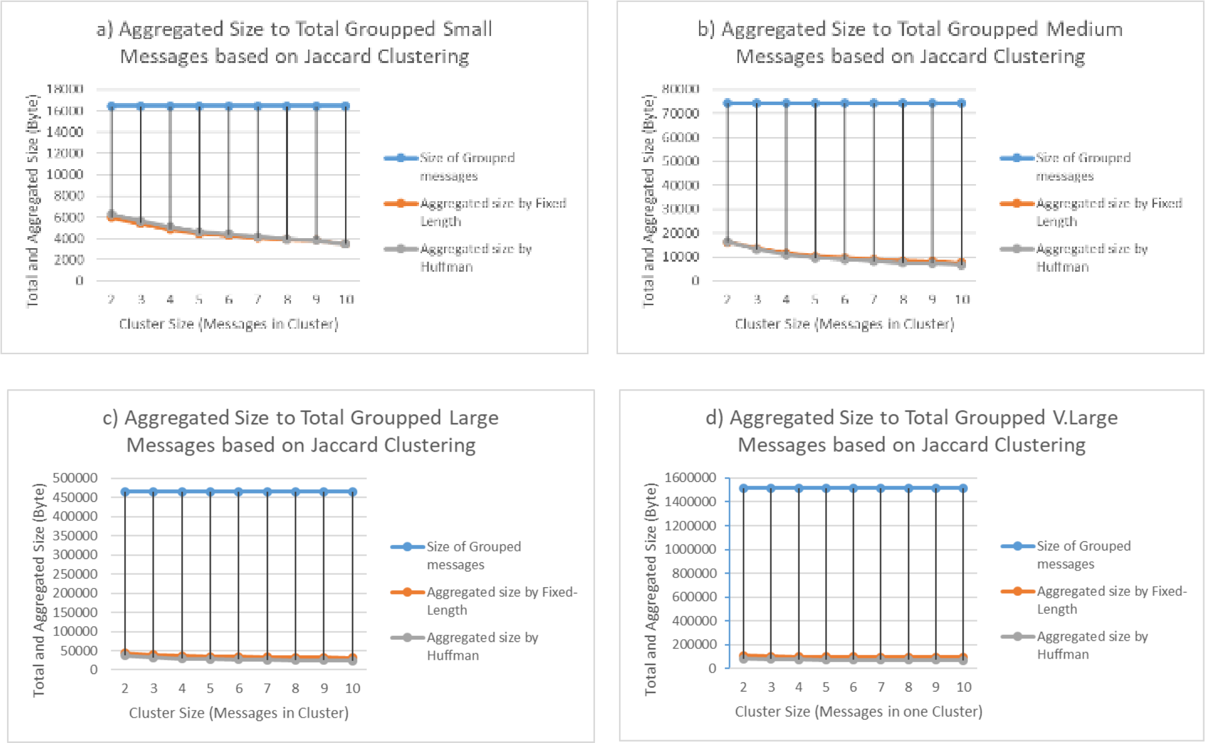


Figure 7: Aggregated Web messages sizes based on Jaccard Clustering

(small, medium, large, and very large) of messages. Messages have been distributed based on their size as potentially small messages like only 140 bytes allocated to small group and potentially large messages like 53KB allocated to very large group. The evaluation strategy for the proposed model is set to test its resultant clusters by Web messages aggregator achieved by [16]. The same aggregator has applied to messages clusters generated by K-Means and PCA combined with K-Means clustering respectively. Aggregator resultant size reduction, compression ratios and clustering processing time are the main metrics to test our proposed model performance. Overall, the proposed model has potentially outperformed other models.

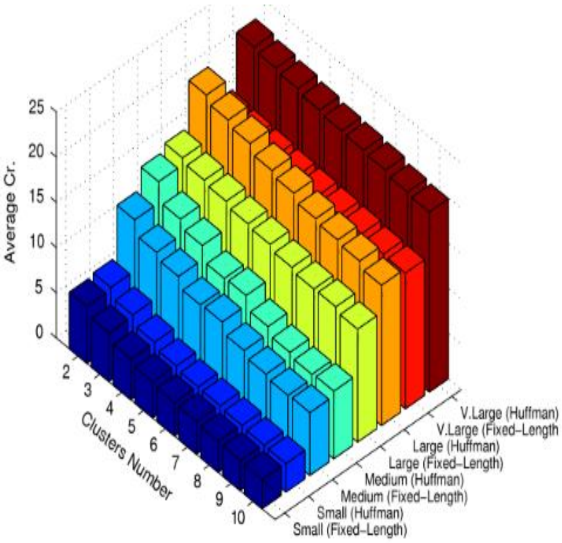


Figure 6: Aggregated Web messages based on K-Means and K-Means + PCA clustering

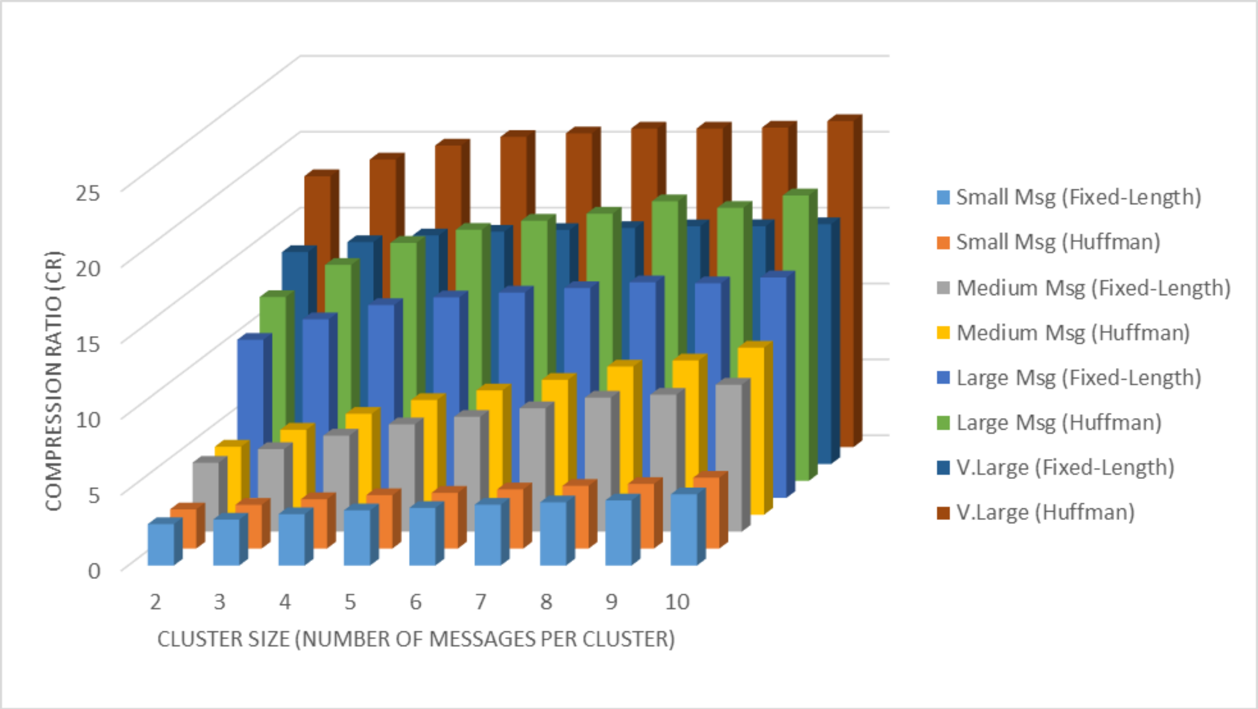


Figure 8: Average Compression Ratios of Aggregated messages based on Jaccard Clustering

Figure 6 shows detailed observation for the Jaccard clustering size reduction in comparison with accumulated size of grouped messages without any size reduction. The results show a tremendous network size reduction that can significantly improve network performance and solve network bottlenecks and congestions. Table 1 provides a summary of the overall average compression ratio and clustering time for all models. Our proposed model has competitive compression ratios with K-Means and PCA combined K-Means as has generally outperformed them. Moreover, Jaccard clustering has potentially outperformed other benchmark models by processing time as it only requires about %32 and %25 of K-Means and PCA combined K-Means processing time respectively. Figures 7 and 8 shows detailed compression ratios for all the involved models and clearly Jaccard clustering has outperformed other techniques. Figure 9 illustrates the processing time in comparison with other benchmark models as the proposed technique has potentially outperformed them.

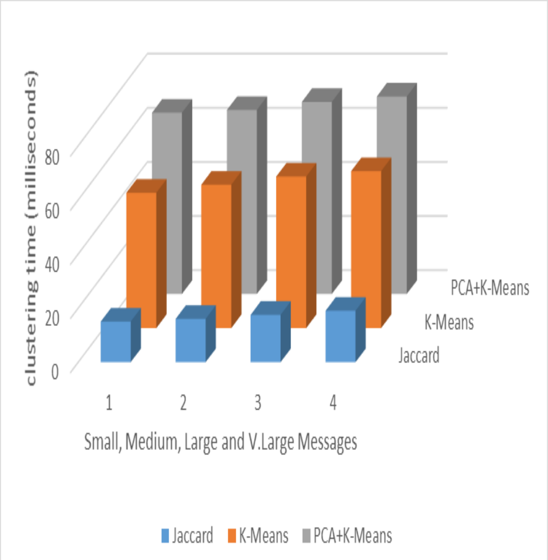


Figure 9: Average Clustering Time for Jaccard, K-Means and PCA+K-Means

5. Conclusions

Web messages have been generated and used significantly over the network to exchange information. Generally, networks suffer from bottlenecks and congestions. Aggregation of Web messages would improve networks potentially. Efficient clustering would be a significant alternative to basic similarity measurements. This paper has proposed Jaccard Coefficient for clustering based on a predefined cluster size. Experiments have shown great achievement by the proposed model in comparison with both K-Means and PCA combined with K-Means as it outperformed them in both resultant compression ratios and clustering time. For future work, we are planning to deploy Jaccard clustering inside a large scale inter-cloud and measure its real-time performance.

Table 1: Average compression ratio and clustering time of K-Means, PCA combined K-Means, and Jaccard based clustering

Cr and Time Consumption	**K-Means**	**PCA+K-Means**	**Jaccard**
Small Messages			
Cr (Fixed-Length)	3.92	3.85	3.75
Cr (Huffman)	3.82	3.71	3.65
Clustering Time (ms)	50.88	65.33	15
Medium Messages			
Cr (Fixed-Length)	6.77	6.8	7.4
Cr (Huffman)	7.72	7.84	8.04
Clustering Time (ms)	52.33	62.89	16
Large Messages			
Cr (Fixed-Length)	12.94	12.82	13.15
Cr (Huffman)	16.02	16.28	16.51
Clustering Time (ms)	54	68.11	17.5
V.Large Messages			
Cr (Fixed-Length)	15.12	15.13	15.23
Cr (Huffman)	20.16	20.25	20.26
Clustering Time (ms)	53.62	70.44	19

6. References

[1] AK, Mishra J. Enhancing the beauty of fractals, ICCIMA'99. In: Proceedings. Third international conference, New Delhi, India, vol.16, issue 4, 23–26 September 1999. p.454–8.

[2] Al-Shammary D, KhalilI, George L. Clustering SOAP web services on internet computing using fast fractals.In:201110 the IEEE international symposiumon network computing and applications (NCA),2011.p.366–71.

[3] C, Yang Z, Peng Z, Hua D, Xiaoxiao H, Zheng W, Junliang C. Development of web- telecom based hybrid services or chest ration and execution middleware over convergence networks.JNetworkComputAppl2010;33:620–30.

[4] C. Werner, C. Buschmann, Compressing SOAP messages by differential encoding, Web services, 2004, IEEE International Conference on Web services, pages 540-547

[5] D. Davis, M. Parashar, Latency performance soap implementations, in: 2nd IEEE/ACM International Symposium on Clust Computing and the Grid, 2002,

[6] Dhiah Al-Shammary, Ibrahim Khal Redundancy-aware SOAP messag compression and aggregation for enhanc performance, Journal of Network ar Computer Applications, Volume 35, Issue 2012, Pages 365-381

[7] Dikaiakos M, Katsaros D, Mehra P, Pallis (Vakali A. Cloud computing :distributed intern computing for it and scientific research .IEE Internet Comput 2009;13:10–3.

[8] G. Yongming, C Dehua, L. Jiajin, Clusterir XML Documents by Combining Content ar Structure, Volume 1, 2008, Pages 583-587

[9] J.C. Hart, Fractal image compression ar recurrent iterated function systems, in: Computer Graphics and Applications, vc 16, IEEE Computer Society Press Lo Alamitos, CA, USA, 1996, pp. 25–33.

[10] J.H. Hwang, M.S. Gu, Clustering xml documents based on the weight of frequent structures, Nov. 2007, pp. 845–849.

[11] M.N. Jain, A.K. Murty, P.J. Flynn, Data clustering: a review, ACM Computing Survey 31 (3) (1999) 264–323.

[12] N. Abu-Ghazaleh, M. Lewis, Differential deserialization for optimized soap performance, in: Supercomputing, 2005, in: Proceedings of the ACM/IEEE SC 2005 Conference, Nov. 2005, p. 21

[13] S. Flesca, G. Manco, E. Masciari, L. Pontieri, A.

Pugliese, Fast detection of xml structural similarity, IEEE Transactions on Knowledge and Data Engineering 17 (2) (2005) 160–175.

[14] V, Makris C, Panagis Y, Sakkopoulos E. Techniques to support web service selection and consumption with QoS characteristics. J Network Comput Appl 2008;31: 108–30.

[15] X. Cui, T. Potok, P. Palathingal, Document clustering using particle swarm optimization, Jun. 2005, pp. 185–191.

[16] Z.T. Khoi, Anh Phan, P. Bertok, Similarity-based soap multicast protocol to reduce bandwidth and latency in web services, IEEE Transactions on Services
Computing 1 (2) (2008) 88–103.

تجميع وتقليل احمال الشبكات باستحداث مصنف رسائل الاكس ام ال باعتماد معامل التشابه لجاكارد

م.د ضياء عيدان جبر الشمري
قسم علوم الحاسوب / كلية علوم الحاسوب وتكنولوجيا المعلومات / جامعة القادسية
d.alshammary@qu.edu.iq

المستخلص:

يقدم هذا البحث مصنف جديد لدعم عمليات تجميع وتقليل احمال الشبكات الخاصة برسائل الاكس ام ال بالاعتماد على معامل تشابه جاكارد. النظام المقترح يعمل على توزيع الرسائل المتشابه بشكل عالي فقط لادخالها لانظمة التجميع. تجميع الرسائل الشبكية اصبح مؤخرا من الحلول الناجعة لمشاكل الشبكة والاختناقات من خلال التخلص من المعلومات المتكررة بالرسائل. تم مقارنة اداء المصنف المقترح مع تقنيات الكي-مينس و تقنية التحليل الجزئي الملحق بتقنية الكي-مينس. اظهرت نتائج المقارنة اداء متميز لمصنف الجاكارد حيث يستهلك المصنف مايقرب بمعدل 32% و 25% مما يستهلكه وقت المعالجة لتقنيات الكي-مينس وتقنية التحليل الجزئي الملحقة بتقنية الكي-مينس بالتتابع. كذلك النتائج اظهرت نسبة ضغط تجاوزت ماانجز مع التقنيات المقارنة.