A Bayesian approach and probabilistic latent variable clustering based web services selection

K. Vaitheki*

Department of Computer Science,
Pondicherry University,
Puducherry, India
Email: vaidehi.balaji@gmail.com
*Corresponding author

G. Zayaraz

Department of Computer Science,
Pondicherry Engineering College,
Puducherry, India
Email: gzayaraz@pec.edu

Abstract: Web services are the product framework to help interoperable machine to machine connection over a system. There is a constant increase in the number of services and processing the large quantity of data over the web requires the exceptional and an improved service selection and classification approach. The requisite to recommend services are grounded on both functional and non-functional requirements. The user keyword extraction using the lexical analyser may give a better extraction than the traditional keyword based search. The lexical analysis process the input character sequences to produce symbol sequences called tokens. The subsequent tokens are then passed on to some other type of formulating and for use as contribution to different assignments, for example, parsers. The tradition of Bayesian system display that is comprehensively utilised for clustering and classifying, is productive for dealing with the non-missing of services. The probabilistic latent variable clustering (PLVC) technique enhanced with the Bayesian classification improves the probabilistic dependencies amongst the clusters and to carry out the clustering task. This may perform better in relations of parameters like precision, recall, and F-measure. The quality of the cluster is foreseeable to be better in terms of purity and entropy for the proposed algorithm.

Keywords: web service; Bayesian network; lexical analysis; probabilistic latent variable clustering; PLVC.

Reference to this paper should be made as follows: Vaitheki, K. and Zayaraz, G. (2017) ‘A Bayesian approach and probabilistic latent variable clustering based web services selection’, *Int. J. Knowledge Engineering and Soft Data Paradigms*, Vol. 6, No. 1, pp.82–93.

Biographical notes: K. Vaitheki is currently working as an Assistant Professor in the Department of Computer Science. She completed her Bachelor of Engineering in Computer Science and Master of Technology in Computer Science with Information Security as specialisation. She won the Pondicherry University Gold medal in Master of Technology. She authored six papers in peer reviewed international journals and co-authored three papers in reviewed journals and have presented papers in three conferences.

G. Zayaraz is currently working as a Professor in Computer Science and Engineering Department at the Pondicherry Engineering College, Puducherry, India. He received his Bachelor, Master and Doctorate degrees in Computer Science and Engineering from the Pondicherry University. He published more than 50 research papers in reputed international journals and conferences. His areas of specialisation include software architecture and information security. He is a reviewer/editorial member for several reputed international journals and conferences and life member of the CSI and ISTE.

This paper is a revised and expanded version of a paper entitled 'A comparative study of K-means clustering based web services selection' presented at International Conference on Communication and Security (ICCS 2016), Puducherry, 17–18 March 2016.

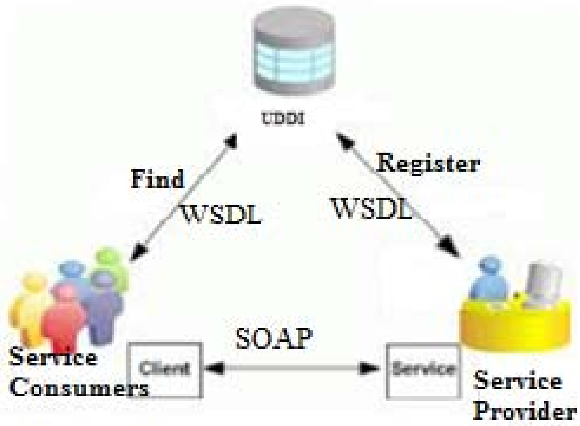
1 Introduction

The interchange of data with each other is mostly required by different software systems and numerous associations do utilise the various programming frameworks for providing services. A web service is a technique for correspondence that empowers two programming structures to exchange this data over the web. Web administrations are the application that utilisations XML to depict a demand and use the uniform resource identifier (URI) to recognise the interface of the application and restricting strategies. Web services are self-portraying application parts that impart utilising open conventions and can be found utilising UDDI. The HTTP and XML are the basis for web services. Web service selection is a requisite procedure for web service configuration so as to first rates the best web service to the client's necessity. The same number of web services are expanded in the web for comparative usefulness, the unparalleled decision of service for the customer necessity is a subtle undertaking for web service administrators.

The product framework that requests for the information is known as a service requester, though the product framework that would technique these demand on request and give the information is known as a *service provider*. A directory called universal description, discovery and integration (UDDI) outlines which software method should be interconnected for which type of data. So when one programming framework needs one specific report/information, it would go to the UDDI and discover which other framework it can contact for accepting that information. The software system categorises and discovers the other system that it should lay a hand on with it would then contact the exacting system by means of a special protocol called simple object access protocol (SOAP). The data request is firstly certified by the service provider system by stating to the WSDL file, and later the request is processed and the data is send beneath the SOAP protocol.

Clustering is an undertaking of collection the task of grouping the set of objects so the items are comparative in a similar gathering. The method reported by Zhanga et al. (2010) at the point when connected to programming segments groups every one of the segments with related element into one class and those with divergent elements into another class. Clustering diminishes the hunt time many-sided quality as all the comparative segments are gathered into one class that is huge and valuable. Clustering is any one special algorithm that has a significant goal then clusters internment the natural structure of the data. Clustering must be observed as the general assignment to be solved. According to George (2013), document clustering or text clustering is one of the main themes in text mining. It alludes to the way toward collecting records with comparable substance or subjects into bunches to enhance both accessibility and unwavering quality of content mining applications such as information retrieval, organising text, summarising document sets, etc.

Figure 1 Components of web services (see online version for colours)



The chief contribution of the proposed work is to analyse the techniques available for clustering of web services

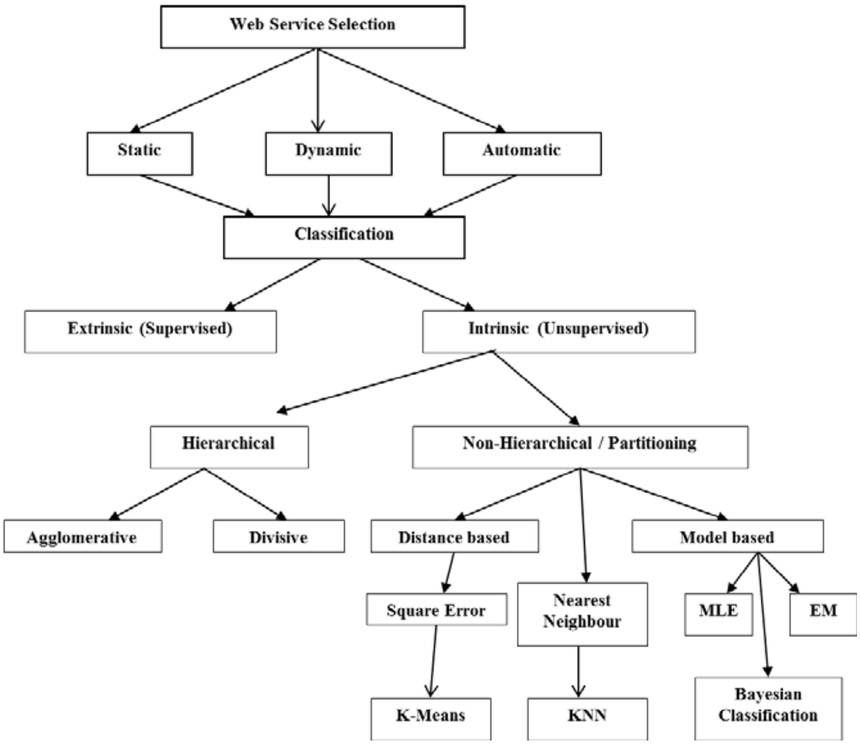
1. A survey report of web services selection based on a clustering technique.
2. The web services selection approach involves lexical analyser and Bayesian network model. The resultant web service selection by classification and clustering would be proficient than the existing approaches.
3. Exploring the viability of classification and clustering algorithm on a extensive variety of unbalanced datasets.
4. The yielding result of the external quality measures like precision, recall, F-measure would probably stay enhanced than the existing approach.
5. The proposed system would be applicable for existing applications, and that it can be used for selection of web services that could function as a precursor for search engines.

Whatever remains of this paper is composed as takes after. Section 2 portrays the related work. In Section 3, a dialog about the related work and proposed work is given. In Section 4 conclusion and a point of view toward research directions, i.e., characterising the diverse issues and the comparing arrangements accessible.

2 Background material

Web service selection is a vital element in service-oriented computing. The most effective method to shrewdly choose suitable web services for the advantages of service users is a key issue in service discovery. The concert of the predictable classification techniques such as decision tree, neural networks, naive Bayes and k-nearest neighbour is hindered by the class inequity problem. In order to lighten the performance deterioration on a large sample datasets, a probabilistic latent variable clustering algorithm is proposed, along with a classification approach by building a Bayesian network model. The following survey shows the problem that is present in the existing clustering approach than the probability based clustering approach. A Bayesian network is a coordinated non-cyclic diagram demonstrates that speaks to restrictive independencies between an arrangement of factors by Al-Masri and Mahmoud (2008). It has two elements: One is a system graphical structure which is a coordinated non-cyclic diagram with the hubs of factors and curves of relations. The other is the restrictive likelihood table connected with every hub in the model diagram. Machine learning procedures can assess the structure and the contingent likelihood table from the preparation information. In view of the Bayesian likelihood derivation, the restrictive likelihood can be evaluated from the measurable information and spread along the connections of the system structure to the objective mark. By locale a limit of certainty, the final likelihood probability value can be utilised as the suggestion for the grouping choice.

Figure 2 Classification of web services selection techniques



3 Related work

3.1 Modified K-means

In sight of Patel and Mehta (2011), modified K-means has addressed two shortcomings

1 Pass number of centroids in apriori and does not lever noise.

2 An indication of cluster analysis with algorithm, pre-processing and normalisation techniques.

The efficiency and productivity of the algorithm proposed by Patel and Mehta (2011) was found to be improved than the standard K-means. It concentrates only on a pre-processing and normalisation. The proposed imparts a lexical analysing scheme over the pre-processing stage and Bayesian classification model that could further circumvent the missing of data. These steps are to be performed earlier even before than the formation of clusters.

3.2 Ontology based computing

Ontology based computing was considered as regular development of existing advances to adapt to data surge. The foundation learning is derivative from the word net as ontology is connected amid pre-processing of archives for report bunching. Relative examination is done between grouping utilising K-means and bunching utilising bisecting K-means. The outcomes of bisecting K-means are established to be better than standard-means clustering technique. The results depend on examination of essentials of clustering algorithm and nature of the report information. The exertion results are based on the root mean squared error (RMSE) parameter (Reuter's top 10 dataset). It is based on the user predicted rating. As the user predicted rating may lead to admittance of malicious user or intended user feedback the proposed work does not strictly rely on the user rating.

3.3 Buckshot document clustering algorithm

The parallel Buckshot report grouping calculation was proposed. This comparing approach is exceedingly taught regarding load adjusting and minimisation of correspondence. Parallel approach was appeared to be adaptable as far as processors productively utilised and number of groups made (Jenson et al., 2002). Load balancing may not be an issue because the offline dataset is regular at the set period of time based on the proposed algorithm. The busy services of online can be selected from the offline dataset. So communication minimisation necessity does not arise.

3.4 Clustering method based on K-means

The K-means based clustering method combine the leading minimum distance algorithm and usual K-means to propose an enhanced K-means clustering algorithm. The value-added K-means technique makes up the deficiencies for the traditional K-means algorithm to state the initial focal point. The improved K-means effectually resolved two drawbacks of the traditional algorithm.

1 Larger dependency to choice the initial focal point.

2 Informal to be confined in the local minimum.

The authors Li and Wu (2012) guaranteed the enhanced K-means is clearly superior to anything the standard K-means in both group accuracy and dependability. Enhanced K-means keeps the high effectiveness of the standard K-means yet in addition raises the speed of union viably by enhancing the method for picking the underlying bunch point of convergence. In the proposed, the cluster focal point is not a random selection wherein based on the conditional probability of the keyword and the latent variable. The transformation of cluster head may not be required.

3.5 Autonomous K-means clustering

In Elomaa and Koivistoinen (2005), autonomous versions of K-means algorithm the shortcomings and deficiencies are identified and overcome. If the user may not need to supply the quantity of bunches for the calculation it would have been considerably richer. The semantics and the probabilistic model were not considered hence the chance of missing documents and services. The proposed work overlook semantics and Bayesian network model to avoid missing of services.

3.6 Hierarchical hesitant fuzzy K-means clustering

The effort by ChenNa and Xia (2014), focused on exploring the clustering techniques for hesitant fuzzy sets based on the K-means clustering and hierarchical clustering. The usual method of modification of K-value is random selection in K-means. This paper claims to fix initial K-value for forming clusters. Such usage significantly lessens the iterative circumstances emerging from choosing the original seeds chaotically utilised as a part of the first K-means. The discussion is only about the fixed K-value i.e., the initial clusters. As it is a fuzzy clustering method, the set of objects may be repeated in different clusters with certain degree of membership. The proposed will overcome the repetition of objects by using the Bayesian networks and latent variable for formation of clusters.

3.7 Heuristic frequent term-based clustering

In this examination, a novel heuristics for grouping news features were proposed by Bora et al. (2012).The versions of frequent term based and frequent noun based clustering algorithms were employed grounded on heuristics. The cluster quality evaluation measures (purity, entropy and F-measure) were found to yield a better performance than the traditional clustering algorithm. The value of the purity, entropy and F-measure in proposed is expected to be better as it has involved a Bayesian classification of services.

3.8 K-means algorithm for market segmentation

Kuo et al. (2002) have equated three clustering methods such as the conventional two stage method, the self-organising feature maps and proposed two stage methods. The two phase strategy is simply the mix of the sorting out component maps and K-means technique. The reproduction comes about demonstrate that the proposed conspire is marginally superior to the ordinary two phase technique as for the rate of misclassification. The above method is found to be slightly better than the conventional two stages. The proposed does not deal with misclassification due to the latent variable and Bayesian model which results in purely domain based cluster.

3.9 Clustering probabilistic semantic approach

Clustering probabilistic semantic approach proposed by Ma et al. (2008) had overcome the problems due to dearth of semantics and great cost of computation. The irrelevant services are eliminated during K-Means approach. After the removal of insignificant services regarding a question, PLSA strategy is connected to the administration dataset with the goal that administration coordinating against the inquiry is done at idea level. The overall results show that our approaches would improve in precision and recall. There is chance of missing data as only two general filtering approaches used by Ma et al. (2008). For cluster formation K-means clustering algorithm was cast-off. Hence, the reformation of cluster head lead to problem. The proposed approach deals with three filtering levels of lexical analysis, Bayesian network and clustering to achieve service selection.

3.10 Constraint-partitioning K-means

George (2013) has discussed two approaches that could be used for clustering namely PCA and cop-K means to condense the dimensionality on excessive dimension dataset. This tactic is very operative in generating specific clusters compared with original cop-K Means algorithm .The author has claimed for precise cluster formation and does not deal with filtering of services or documents. The proposed approach deals with three levels of filtering to eliminate irrelevant services and form a precise cluster.

3.11 K-means clustering with PSO

Two approaches for clustering K-means and particle swarm optimisation (PSO).The algorithm is evaluated on the basis of accuracy, execution time ,quantisation error, inter intra cluster distance (Saini and Kaur, 2014). The variation in PSO algorithm and its sequential hybridisation with K-means algorithm if done more efficiently, the execution time can be reduced. The general quality measures are only considered wherein the quality of cluster has not been addressed.

4 Discussion

The learning algorithm with respect to K-means requires apriori description about the number of randomly chosen cluster centres that may not yield the bountiful result.

Table 1 Comparative analysis of existing system

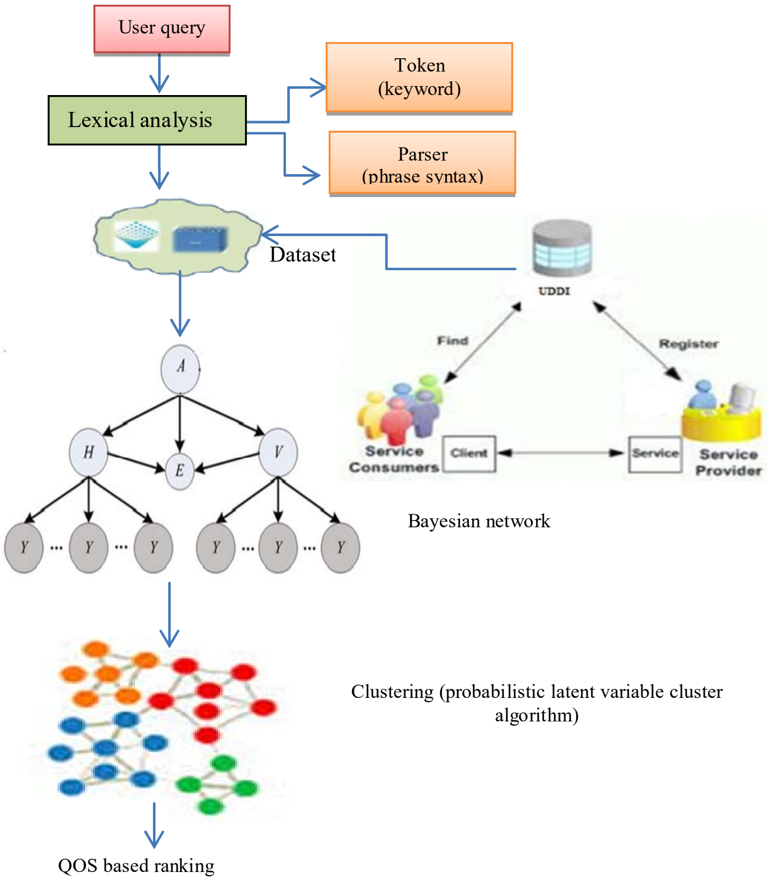
Name of researchers	*Cluster quality*						*Tool/ dataset*
	Precision	*Recall*	*F-measure*	*Purity*	*Entropy*	*Distance*	
Wu et al.	✓	✓	x	x	x	Normalised google distance (NGD)	Dataset (seekda)
Vijayan	✓	✓	x	x	x	x	Word vector tool
ChenNa and Xia	x	x	x	x	x	Cluster formation-centroid and distance was computed	Hesitant fuzzy set (HFS)
Saini and Kaur	x	x	x	x	x	Inter and intra Cluster distance	x
Vamsi et al.	x	x	x	x	x	Cosine similarity measure	x
Li and Wu	✓	x	x	x	x	Euclidean distance	x
Bora et al.	✓	✓	✓	✓	✓	x	x
Ma et al.	✓	✓	x	x	x	x	x
Kuo et al.	x	x	x	x	x	Neighbourhood distance	x

The cluster formation is to be based on the fixing up some initial boundary constraints on the K values of K-means or the integration of K-means with other optimisation approaches that aims to reduce the cross validation errors. The K-means do not effort well with non-globular clusters and more than that with clusters of different size and density. The different initial partitions may result in different final clusters. The estimation of quality of the clusters becomes a difficult task. It is impotent to handle noise and outliers. It is applicable only when mean is defined and abruptly fails for the categorical data. The violation of the rules of the k-means were the variances of all attribute are always spherical and all the variables have the same variances may cause a failure. The Table 1 shows the several diverse techniques are available for selection of Web services based on clustering but those are integrated with K-means algorithm rather than model based probabilistic clustering approach.

5 Proposed architecture and algorithm

The user input query is lexically skimmed for the group of characters that have collective meaning , token (keyword) and lexemes (identifier) being an actual instant for token. The lexical analyser peruses in a surge of characters, distinguishes the lexemes in the stream, and sorts them into tokens. The subsequent user queries are similar to the past input and matches with the token the recurrence of the phases are clogged. The lexical analyser streamlines the design and improves the efficiency and portability. The keyword extracted from the input user query through lexical analyser is complemented with the QWS (Al-Masri and Mahmoud, 2007) and the Titan dataset to extract the related services. The extracted list of services leads to the construction of the Bayesian network, the graphical model capable of displaying relationships clearly and intuitively. The Bayesian model is responsible for non-missing of data. The depth first search is involved to retrieve the services from the Bayesian network model and proceeds with the cluster formation based on the probabilistic latent variable clustering algorithm. This is probable to result in a cluster formation with the conceivable quality with highly relevant services.

Figure 3 Web services selection based on Bayesian classifier and PLVC (see online version for colours)



Algorithm 1 Bayesian classification

```
Input: LA and Dataset (List of selected web services based on domain)
            Web services $WS_i$ ……. $WS_n$, I, Root service $Rs_i$, Graph $G_s$, Vertices $V_i$, and Edges
            Ei, n
Output: non missing of web services (graph constructed denoting web services based on
domain)
Begin
      Select ordering of web services $WS_1$, $Ws_2$ …… $WS_n$
      For i = 1 to n
Begin
      Add $WS_i$ to RS
      Select Rs from $WS_i$ ….. $WS_{i-1}$ such that
      Rs ($WS_i$/Rootservice ($WS_i$)) = Rs ($WS_i$/$WS_i$ ….. $WS_{i-1}$)
End
For each service $WS_i$, add to Rs, the set of vertices $V_i$ such that
   For each $V_i$, edge =$V_i \rightarrow V_j$ in the DAG of $G_i$
     For each directed edge in $G_i$
       Begin
         If (i → j = directed edge) && if ($WS_i$ and $WS_j$ = Rs in G1)
        Begin
        List out the $WS_i$
        Generate the total order on the nodes from this DAG.
```

Algorithm 2 Cluster formation (PLVC – probabilistic latent variable clustering)

```
Input: keyword, Services gained out of Bayesian Classification
Output: Resultant Cluster, Retrieval of all matching services based on the latent variable
Begin
For each domain, fetch the keyword from LA and Latent variable from the BN.
P (d/$W_k$, $L_v$) = Outcomes of relevant services
If ($Ws_i \neq Ws_{i+1}$)
Begin
List = $Ws_i$
End
Else
Remove duplicate services;
End
```

```
If ((keyword $\neq$ NULL) && (LV $\neq$ NULL))
Begin n
$P(Lv \rightarrow Ws_i) = \sum_{i=1} P(Ws_i/Ws_n)$
$P \rightarrow Ws_i = Ws_i/Ws_n$
Return max (P ($Ws_i$))
Head = P ($Ws_i$)
End
If Head < P (Wsi)
Neighbour = P(Wsi)
While (P ($Ws_i$) $\neq$ NULL)
Neighbour = P ($Ws_i$)
end
else
list1 = useless service
end
```

6 Conclusions and future work

The selection of web services is achieved through cluster formation which may result in the listing of relevant services. The Bayesian network has been defined and as a result cluster is formed with the user domain as the cluster head. A few parameters are to be verified with respect to the cluster formation. The lexical analysis phase for extracting the tokens would enrich the search better than the traditional keyword based search. The Bayesian network model approach would lead to non-missing of data and construction of a sequence for verification of domain based web services. The probability based latent variable analysis fetches the highly matching web services and hence forms the cluster. The external evaluation measures for clustering web services is to be evaluated based on the term frequency, precision, recall, F-measure, similarity measure (normalised Googelian distance) and the quality of the cluster (purity and entropy).

References

Al-Masri, E. and Mahmoud, Q.H. (2007) ‘QoS based discovery and ranking of web services’, *IEEE 16th International Conference on Computer Communications and Networks* (*ICCCN*), pp.529–534.

Al-Masri, E. and Mahmoud, Q.H. (2008) ‘Investing web services on the world wide web’, *The 17th International Conference on World Wide Web*, Beijing, pp.795–804.

Bora, N.N., Mishra, B.S.P. and Dehuri, S. (2012) ‘Heuristic frequent term-based clustering of news headlines’, *Second International Conference on Communication, Computing & Security*, Vol. 6, pp.436–443.

ChenNa, X.Z-S., and Xia, M-M. (2014) ‘Hierarchical hesitant fuzzy K-means clustering algorithm’, *Journal of Applied Mathematics in Chinese*, Vol. 29, No. 1, pp.1–17.

Elomaa, T. and Koivistoinen, H. (2005) 'On autonomous k-Means clustering', *International Symposium on Methodologies for Intelligent Systems* (*ISMIS 2005*), LNCS, Vol. 3488, pp.228–236.

George, A. (2013) 'Efficient high dimension data clustering using constraint – partitioning K-means algorithm', *The International Arab Journal of Information Technology*, Vol. 10, No. 5, pp.467–476.

Jenson, E.C., Bietzel, S.M., Pilotto, A.J., Goharian, N. and Fruder, O. (2002) 'Parallelizing the buckshot algorithm for efficient document clustering', *Proceedings of the eleventh international Conference on Information and Knowledge Management*, ACM, pp.684–686.

Kuo, R.J., Ho, L.M. and Hu, C.M. (2002) 'Integration of self-organizing feature map and K-means algorithm for market segmentation', *Journal of Computations and Operations Research*, Vol. 29, No. 11, pp.1475–1493.

Li, Y. and Wu, H. (2012) 'A clustering method based on K-means algorithm', *Physics Procedia*, Vol. 25, No. 2012, pp.1104–1109.

Ma, J., Zhang, Y. and He, J. (2008) 'Efficiently finding web services using a clustering semantic approach', *Proceedings of the International Workshop on Context Enabled Source and Service Selection* (*CSSSIA*).

Patel, V.R. and Mehta, R.G. (2011) 'Modified K-means clustering algorithm', *Proceedings of the First International Conference on Computational Intelligence and Information Technology*, Springer, pp.307–312.

Saini, G. and Kaur, H. (2014) 'A novel approach towards K-means clustering algorithm with PSO', *International Journal of Computer Science and Information Technology*, Vol. 5, No. 4, pp.5978–5986.

Zhanga, W., Yoshida, T., Tang, X. and Wang, Q. (2010) 'Text clustering using frequent itemsets, knowledge-based systems', *Journal of Knowledge Based Systems*, Vol. 23, No. 5, pp.379–388.