Web Services Clustering Approaches: A Review

Abdelmoniem Helmy¹, Akram I. Salah², Mervat H. Geith³

¹ Information Systems dept., Institute of Statistical Studies & Research, Cairo University.
² Computer Science dept., Institute of Statistical Studies & Research, Cairo University.
³ Computer Science dept., Faculty of Computers & Information Science, Cairo University.

Abstract :- Web services clustering is a process of dividing data objects into several classes or clusters to make the similarity between different clusters minimal and the similarity between objects within one cluster maximal. Clustering is considered the most important non-supervised learning technique that can organize the web services in such a way that the similar web services in a cluster provide similar matches with user's request. With the rapid growth of available web services developed by different organizations, clustering of web services is required for conveniently managing web services selection, discovery, and composition and QoS prediction. As the performance of Web services clustering relies closely on services representation, the similarity measure, and the clustering algorithm, in this paper we will review methods, techniques, and approaches for web services clustering and elaborate how much different approaches contribute to an efficient and effective service discovery.

Keywords: Web Services, Web Services Clustering, Clustering Approaches

I. INTRODUCTION

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). With the growing number of services in the repositories becomes challenge for quickly finding of web services. For this reason, clustering related services in specified cluster is required. This paper review variations of algorithms, general approaches and modern approaches for web services clustering. The paper is organized as follow; the following section highlight current status of knowledge about this topic. Section 3 compare between two main algorithms for clustering, namely partition based and model based clustering. Section 4 contrasts among variations of traditional approaches for web services clustering. Specifically, partitioning versus hierarchal, syntactic versus semantic, and functional versus nonfunctional clustering approaches. In section 5, modern approaches for web services clustering are introduced. These approaches provide an intelligent clustering methods using biologicallyinspired methods, taxonomic clustering, context based clustering, and fuzzy clustering to avoid poor accuracy introduced by classical approaches. The conclusion for this review is presented in section 6.

II. CURRENT STATUS OF KNOWLEDGE

Web service discovery is becoming difficult task because of increasing Web services available on the Internet. As seeking for efficient web service discovery is main challenge for researchers, research in cluster analysis of web services has recently gained much attention. This is due to the popularity of web services and the potential benefits that can be achieved from cluster analysis of web services like reducing the search space of a service search task. Web services analysis is the basis of using Web services discovery effectively and efficiently. The first step in such analysis of Web services is to categorize different services, which may be offered by different service providers, based on their functionalities. The main motivation behind clustering the service in a directory is to re-organize the linear structure into a tree structure; hence efficient tree-based search algorithm can be applied to service discovery [16]. many researches contributed to the improvement of service discovery using different either partitioning or hieratical clustering methods. The clustering approaches also differ of either using semantic description (OWL-S) or non-semantic description (WSDL) files for the web services to extract valuable information about the service. In addition, for who adopted WSDL-based clustering, they also differ in feature selection among service content, messages, types, ports, binding, context, host name, and service name. Also other researches augmented the tags associated to web services either by service publishers or consumers as an important source of information. In all cases, WSDL, OWL-S, and tags, vector space model (VSM) was adopted as representation model for web service, and two main similarity measures in the literature namely cosine similarity with term frequency inverse document frequency (TF/IDF) and Normalized Google Distance (NGD).

While, some research focus on text similarity between web services, others highlight the functional

similarity using service input, output, precondition and effect as base for similarity. In spite of that, all researches has common framework that cluster web services before matching user query to match the query with few clusters instead of many individual service, some of work adopt offline clustering while others prefer online clustering. In this context, while most research re-execute the clustering process every time for discovering, some researchers adopt keeping cluster index to be updated for new added services. Moreover, while most researches cluster based on functional attributes, we will find some literatures go one extra step to cluster service based on non-functional quality attributes i.e. quality of services (QoS) attributes. As consequence for last approach, fuzzy techniques was introduced as the maestro of clustering process as the quality attributes cannot be decided as simple Boolean existence rule. Also, while most researchers cluster services for aim of similarity, the others adopted clustering for aim of composition. Finally, while most researchers used classical clustering techniques such as partitioning k-means and agglomerative hierarchal clustering, some researchers adopted modern clustering techniques such as practical swarm optimization (PSO) and Ant-Colony algorithms

Clustering algorithms can generally be divided into five categories; hierarchy based approach, division based approach, density based approach, grid based approach and model based approach. Each method has its own advantages and disadvantages [1]. An entire collection of clusters is commonly referred to as a clustering, and it can be divided into threemain categories: partition based, model based, and hierarchical clustering. In the next section will explain two different techniques used for service clustering namely, partition based clustering and model based clustering.

III. WEB SERVICES CLUSTERING ALGORITHMS

3.1 Partition based Clustering

Sometimes referred to as objective function based clustering [2]. These algorithms minimize a given clustering criterion by iteratively relocating data points between clusters until a (locally) optimal partition is attained [3]. While the algorithmic setup is quite appealing and convincing (the optimization problem could be well formalized), one is never sure what type of structure to expect and hence what should be the most suitable form of the objective function. Typically, in this category of clustering techniques, we predefine the number of clusters and proceed with the optimization of an objective function [4]. The algorithms from this category are: K-means [5], [6] and Self Organizing Feature Maps (SOM) [7]. These algorithms were used with the objective to create groups of queries which have in common certain features such as the topic or the user's goal. In particular, the former model, i.e., K-means was used to facilitate and speed up the manual classification of different groups of queries by creating a context to each query, and to give sense to some queries which, from their terms, are incomprehensible, hence difficult to classify.

3.2 Model based clustering

In model based clustering, we assume a certain probabilistic model of the data and then estimate its parameters [3]. In this case, it refers to a so called mixture density model where we assume that the data are a result of a mixture of c sources of data. Each of these sources is treated as a potential cluster [8]. The algorithm from this category is Probabilistic Latent Semantic Analysis [9] [10]. The aim of using this algorithm was to discover the hidden relationships that underlie the queries, and thus, to determine characteristics of the queries such as the topic or the motivation of the user to pose a query.

Different examination applied semi-supervised clustering in web services composition [15]. A semisupervised clustering algorithm uses a large amount of unlabeled data, supporting the supervised learning process to improve clustering results. Semi-supervised learning can be divided into semi-supervised clustering and unsupervised clustering. The particular semi-supervised clustering algorithm used in this approach clusters data based on tags and constraints. The tag data indicate an instance of a particular category. Two instances should belong to the same cluster. The algorithm is modified to meet the specific constraints of a service. This application of semi-supervised clustering has been applied to web services composition. This particular examination demonstrates that the application of semi-supervised clustering in web services composition shows potential for improving results. Both the unsupervised and semi-supervised approaches have produced efficient clustering results.

IV. GENERAL APPROACHES FOR CLUSTERING

The problem of web services clustering can be approached by using the classical partitioning and hierarchical clustering techniques. A comparison between the partitioning and hierarchical techniques is presented in [11] and can be summarized as follows; partitioning clustering as opposed to hierarchical clustering avoids the problem of overlapping clusters, while hierarchical clustering as opposed to partitioning clustering does not require to know in advance the number of clusters to be discovered. Two methods for service clustering

using a classical technique are presented in [12] [13]. The method proposed in [13] relies on the hierarchical agglomerative technique to identify service clusters based on a given query and by considering the services' syntactic descriptions as clustering criteria. On the other hand, the method proposed in [12] applies the hierarchical agglomerative technique to cluster services based on their semantic descriptions.

4.2 Syntactic vs. Semantic Clustering

In addition to improvements in classification, the process of matching services to requests could also utilize better techniques for analyzing web services. One mode of web services analysis uses syntactic structure. Some research has sought to improve the discovery of web services with search engines by proposing a new approach to clustering WSDL documents into functionally-similar groups before answering discovery requests [17]. This approach mines the WSDL to extract features that describe the semantics and behavior of the web service, which reveal the functionality of the service. By integrating these features together, the approach then clusters web services into functionally-similar groups.

The work in [18] introduce WordNet lexical database to work with the Vector Space Model (VSM) feature vectors for web services' patterns representation. Besides, they employ Self Organizing Map (SOM) neural network approach in clustering method. The good representation of web services' patterns produced by the lexical database has helped to improve the clustering process. This work however only considers functional descriptions in performing the clustering.In their work [26], Zhao and Chen introduced A service automatic classificationapproach based on WordNet by combining text mining, semantictechnology and machine learning. In this work, their techniques of text mining rely only on text description of services so that it can classify differenttype services, such as WSDL Web Service, RESTfulWeb Service and traditionalnetwork based software component service.

4.3 Functional vs. Non-Functional Clustering

There may be multiple service providers who offer the same functionalities defined in a service interface. Determining and choosing the best service becomes important for service requesters. The information in WSDL descriptions is not sufficient for ranking best services. Non-functional properties including specification of the cost, performance, security, and trustiness of a service are introduced for measuring the Quality of Services (QoS). There are many aspects of QoS that can be organized into categories with a set of quantifiable parameters [20]. The "best" service may have different meanings for different requesters. One may prefer security over cost while the other may prefer lower cost over performance. Measurements of these non-functional properties can be achieved using statistical analysis, data mining, and text mining technologies. It is normally done by a third-party through the collection of subjective evaluations from requesters. This information dynamically changes over time [22].

The work in [23] describes a clustering based search approach for Web services, which is implemented in their search engine, Woggle. Their tool extracts a set of semantic concepts using natural language descriptions that have to be provided by the Web service in the repository. This work however only considers functional descriptions in performing the clustering.[21] propose a clustering approach to improve the search process as they produce a clustering semantic algorithm, which has the capability toeliminate the return of irrelevant services from users' queries by using probabilistic Latent Semantic Analysis (PLSA).[25] argues that web services orchestration only takes into consideration the descriptions of what the services areand how they are provided. It ignores the other important considerations that describe when and where the services areserved.

The work in [28] investigates the clustering of web services' users based on their interests. This information is analyzedand compared with web services' characteristics to discover the matching services. Using Quality of Service(QoS) that defines the nonfunctional requirements of a service such as response time, price, and availability, the service is evaluated in order to confirm whether it meets the users' expectation or not.[27] argue that web services matching processes are slow as a result of complex semantic calculations. Thus, they proposea clustering method that is based on user preferences and ontology. [29] suggest semi-supervised clustering algorithm for web services composition. The algorithm performs the clustering based on tagsand constraints, and uses a huge number of unlabeled data in order to support the supervised learning procedure. The authors in [35] are criticizing that Most systems assume that users could formulate their QoS requirements easily and are accurately using the provided query languages. Hence, they emphasize more on the user-centered design of the service selection system by providing a more expressive and flexible way for non-expert users to define their QoS queries.

V. MODERN APPROACHES FOR WEB SERVICES CLUSTERING 5.1 Biologically-Inspired Clustering

Besides the classical techniques, researchers consider the use of biologicallyinspired methods for clustering [11], [30], [31], [32], and [33]. In [11] it was stated that the clustering method based on Particle Swarm Optimization is better than partitioning clustering as it avoids the problem of local optima stagnation. The advantages of ant based clustering compared to the classical clustering methods are presented in [31]. According to [31] ant based clustering does not require to know in advance the number of clusters and the obtained clusters have a higher quality. Research literature reports a classical [31] and a hybrid [33] ant based clustering technique. The hybrid technique, Tree Traversing Ant (TTA), combines features of ant based clustering techniques. In the case of service clustering, the authorsin [14] have identified a method based on TTA that considers the services' syntactic descriptions as clusteringciteria.

In their work [19], they propose a semantic service clustering method which adapts the ant based clustering algorithmpresented in [31]. This method clusters the services based on their semantic descriptions, as opposed to [14] and[13] which consider only the syntactic descriptions. By considering the semantic descriptions, the clustering process isimproved, as it enables reasoning which leads to refined results. By applying an ant-inspired algorithm for service clusteringit is not necessary to specify in advance the number of clusters to be created and also the problem of overlappingclusters is avoided. In addition, to refine the clustering results [19] employ a method that allows the isolated services to migrate to the most suitable cluster.

Other research [36] proposes the employment of biologically inspired method in semantic web services clustering in order to produce more efficient discovery process. The method evaluates the similarity of two services by comparing the degreeof match of their semantic descriptions. In the other work [37], Particle Swarm Optimization is proposed as a method for web services clustering. The method applies the similar concept to cluster services, which is by evaluating the similarity of services' semantic descriptions. Moreover, the work also defines a set of metrics which is able to perform thissimilarity evaluation.

5.2 Taxonomic Clustering

The authors in [34] propose the use of taxonomic clusteringalgorithm that groups web services based on their functional similarity. [38] investigates web services' functionalityanalysis which utilizes clustering method in its process. Through this clustering method, service taxonomy's hierarchyis created. Any web services will be labeled accordingly after they went through the clustering process. Any new webservices will be compared with the clustered ones and then assigned with the matched labels.Gao, Stucky, and Liu [40] propose an intelligent web services clustering method through combination of humanknowledge and artificial intelligence (AI) technique to generate a taxonomy structure. In the process, web services aretransformed into standard vector format through WSDL document.

5.4Fuzzy Clustering

A different angle of web services clustering leads to a proposal of fuzzy clustering of web services based on quality of service. Clustering is essential to ensuring the efficiency of web service discovery, selection and recommendation. Fuzzy clustering could be used to guide requestors to suitable services. A technique for clustering must have the capability to consider all possible parameter matches and include multiple criteria in the process of parameter matching.

The work [24] proposes a technique that meets these requirements and shows better results with less irregularity. It provides an explanation of how web services' quality of service data can be clustered fuzzily using the unsupervised method of FCM. Additionally, the generated fuzzy-clustering results can be used during contract negotiation and services selection. This benefit aids requestors who have limited technical knowledge of different web services. By applying the processes proposed by this work, the fuzzyclustering results can be updated based on the latest quality of service data. The results of the clustering can also contribute positively to the development of fuzzy based web services' applications. Clustered quality of service data can generate fuzzy inference better than expert knowledge, which suffers from more inaccuracies and the unavailability of technologically-proficient people to provide input. The authors also argue that clustering is essential to producing efficient service discovery, selection, and recommendation. It then proposed the clustering of web services based on their QoS, which has never been implemented before.

The fuzzy clustering of web services could help requestors who have limited technical knowledge about the web services to understand what the realistic quality of service is. They could subscribe to services that give the best value for their money. The clustering could be used as a reference for requestors in the process of negotiating and specifying service requirements. This kind of clustering also contributes to the development of web services based on this fuzzy clustering. The automatic generation of inference components such as the proposed QoS clustering could provide an alternative to expert knowledge that requires less time, suffers from less accuracy, and has wider availability. By prioritizing the requestor, this method could lead to more

productive service discovery. Requestors could find services that fit their specific requirements and understand the actual quality of each service before purchasing the service. Gholamzadeh and Taghiyareh [39] propose a fuzzy semantic clustering algorithm that can perform the calculation of semantic similarity among web services. The algorithm is based on ontology concept that enables the grouping of web services into semantically meaningful clusters. The algorithm utilizes the functionality and interface information that are contained in WSDL document. They include name and text descriptions, annotations of ontology files or elements, operation descriptions, and input and output descriptions.

VI. 6.CONCLUSION

Cluster analysis for web services based on functional similarities would greatly improve the ability of web services search engines to retrieve the most relevant web services in response to user requests. Clustering would link services that possess similar functions to accurately find services for a particular requested function. The main goal of clustering is to group objects in a way that objects in one cluster have high similarity to one another, and are very different from objects in another cluster.

In order to effectively improve the service discovery process by changing the search space, clustering techniques can be used to group similar services based on their functionality which improves the search engine retrieval. This approach provides better performance in terms of service discovery efficiency and effectiveness. The increasing number of web services has also caused bad performance and low efficiency of discovery process. The problem of web services clustering can be approached by using the classical partitioning and hierarchical clustering techniques. The distance similarity between web services can be derived from either service text description or functionality. The matching technique can be either syntactically or semantically. Grouping of similar web services can be according service capability or quality attributes. In all ways, it is noted that classical clustering approaches suffer from some degree of poor accuracy. Hence,some new researches consider more intelligent techniques that make use of biologicallyinspired methods such as ant based, and Particle Swarm Optimization (PSO). Taxonomic clustering for services based on its functionality is also introduced. Context based clustering that considers web services context was came to replace the link or linkcontent based clustering. Finally, fuzzy clustering based on quality of service is essential to ensuring the efficiency of web service discovery, selection and recommendation.

VII. REFERENCES

- [1] Gao, H., Wang, S., Sun, L., & Nian, F. (2014). Hierarchical Clustering based Web Service Discovery. In Service Science and Knowledge Innovation (pp. 281291). Springer Berlin Heidelberg
- [2] Ayramo, S. & Karkkainen, T. (2006) Introduction to partitioning based cluster analysis methods with a robust example, Reports of the Department of Mathematical Information Technology; Series C: Software and Computational Engineering, C1, 136
- [3] Krzysztof, J. C., Roman, W. S., & Kurgan, L. (2007). Data Mining: A Knowledge Discovery Approach.
- [4] CalderonBenavides, L. (2011). Unsupervised Identification of the User's Query Intent in Web Search. Universitat Pompeu Fabra.
- [5] MacKay, D. J. (2003). Information theory, inference and learning algorithms. Cambridge university press.
- [6] Steinbach, M., Karypis, G., & Kumar, V. (2000, August). A comparison of document clustering techniques. In KDD workshop on text mining (Vol. 400, No. 1, pp. 525526).
- [7] Kohonen, T. (2012). Selforganization and associative memory (Vol. 8). Springer.
- [8] Cios, K. J., Swiniarski, R. W., Pedrycz, W., & Kurgan, L. A. (2007, January). Unsupervised learning: Clustering. In Data Mining (pp. 257288). Springer US.
- [9] Hofmann, T. (1999, July). Probabilistic latent semantic analysis. InProceedings of the Fifteenth conference on Uncertainty in artificial intelligence (pp. 289296). Morgan Kaufmann Publishers Inc.
- [10] Jin, X., Zhou, Y., & Mobasher, B. (2004, August). Web usage mining based on probabilistic latent semantic analysis. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 197205). ACM.
- [11] Abraham, A., Das, S., & Roy, S. (2008). Swarm intelligence algorithms for data clustering. In Soft Computing for Knowledge Discovery and Data Mining (pp. 279313). Springer US.
- [12] Nayak, R., & Lee, B. (2007). Web service discovery with additional semantics and clustering.
- [13] Platzer, C., Rosenberg, F., & Dustdar, S. (2009). Web service clustering using multidimensional angles as proximity measures. TOIT ACM Transactions on Internet Technology ACM Trans. Internet Technol., 126.
- [14] Liu, W., & Wong, W. (2009). Web service clustering using text mining techniques. International Journal of AgentOriented Software Engineering, 3(1), 626.
- [15] Lei, Y., Zhili, W., Luoming, M., & Xuesong, Q. (2014). Clustering and Recommendation for Semantic

Web Service in Time Series. KSII Transactions on Internet and Information Systems (TIIS), 8(8), 2743-2762.

- [16] Zheng, Q., and Wang, Y. (2012). "The Application of Semi-Supervised Clustering in Web Services Composition." Advances in Computer Science and Information Engineering 2(169), 683-88. Print.
- [17] Elgazzar, K., Hassan, A. E. & Martin, P. (2010). "Clustering WSDL to Bootstrap Discovery of Web Services." International Conference on Web Services, 147-54. Print.
- [18] Chen, L., Yang, G., Zhang, Y., & Chen, Z. (2010, December). Web services clustering using SOM based on kernel cosine similarity measure. In Information Science and Engineering (ICISE), 2010 2nd International Conference on (pp. 846-850). IEEE.
- [19] Pop, C., Chifu, V., Salomie, I., Dinsoreanu, M., David, T., & Acretoaie, V. (2010). Semantic Web Service Clustering for Efficient Discovery Using an Ant based Method. Studies in Computational Intelligence Intelligent Distributed Computing IV, 2333.
- [20] Ran, S. (2003). A model for web services discovery with QoS. ACM Sigecom exchanges, 4(1), 1-10.
- [21] Zhang, R., Zettsu, K., Nakanishi, T., Kidawara, Y., & Kiyoki, (2009, October). Context based web service clustering. In Semantics, Knowledge and Grid, 2009. SKG 2009. Fifth International Conference on (pp. 192199). IEEE.
- [22] Xiang, X. (2007). Service-oriented architecture for integration of bioinformatic data and applications (Doctoral dissertation, University of Notre Dame).
- [23] Dong, X., Halevy, A., Madhavan, J., Nemes, E., & Zhang, J. (2004, August). Similarity search for web services. In Proceedings of the Thirtieth international conference on Very large data basesVolume 30 (pp. 372383). VLDB Endowment
- [24] Hasan, M. H., Jaafar, J. & Hassan, M. F. (2014). "Fuzzy based Clustering of Web Services' Quality of Service: A Review." Journal of Communications 9(1), 8190. Print.
- [25] Hao, Y., Junliang, C., Xiangwu, M., & Bingyu, Q. (2007). Dynamically traveling web service clustering based on spatial and temporal aspects. In Advances in Conceptual Modeling"Foundations and Applications (pp. 348357). Springer Berlin Heidelberg.
- [26] Zhao, H., & Chen, Q. (2011). "An Automatic Service Classification Approach." Advances in Intelligent and Soft Computing 123, 53140. Print.
- [27] Li, M., & Yang, Y. (2012, August). Efficient clustering index for semantic Web service based on user preference. In Computer Science and Information Processing (CSIP), 2012 International Conference on (pp. 291294). IEEE.
- [28] Zheng, K., Xiong, H., Cui, Y., Chen, J., & Han, L. (2012, April). User clustering based web service discovery. In Internet Computing for Science and Engineering (ICICSE), 2012 Sixth International Conference on (pp. 276279). IEEE.
- [29] Zheng, Q., and Wang, Y. (2012). "The Application of Semisupervised Clustering in Web Services Composition." Advances in Computer Science and Information Engineering 2(169), 68388. Print.
- [30] Charles, A., Menezes, R., & Tolksdorf, R. (2004, April). On the implementation of SwarmLinda. In Proceedings of the 42nd annual Southeast regional conference (pp. 297298). ACM.
- [31] Handl, J., Knowles, J., & Dorigo, M. (2006). Ant based clustering and topographic mapping. Artificial life, 12(1), 3562.
- [32] Marinakis, Y., Marinaki, M., & Matsatsinis, N. (2008). A hybrid clustering algorithm based on honey bees mating optimization and greedy randomized adaptive search procedure. In Learning and Intelligent Optimization (pp. 138152). Springer Berlin Heidelberg.
- [33] Wong, W., Liu, W., & Bennamoun, M. (2007). Treetraversing ant algorithm for term clustering based on featureless similarities. Data Mining and Knowledge Discovery, 15(3), 349381.
- [34] Ma, J., Zhang, Y., & He, J. (2008, April). Efficiently finding web services using a clustering semantic approach. In Proceedings of the 2008 international workshop on Context enabled source and service selection, integration and adaptation: organized with the 17th International World Wide Web Conference (WWW 2008) (p. 5). ACM.
- [35] Mobedpour, D., & Ding, C. (2013). User-centered design of a QoS based web service selection system. Service Oriented Computing and Applications, 7(2), 117127.
- [36] Pop, C. B., Chifu, V. R., Salomie, I., Dinsoreanu, M., David, T., Acretoaie, V., ... & Oprisa, C. (2012). Biologically-inspired clustering of semantic Web services. Birds or ants intelligence?. Concurrency and Computation: Practice and Experience, 24(6), 619633.
- [37] Nagy, A., Oprisa, C., Salomie, I., Pop, C. B., Chifu, V. R., & Dinsoreanu, M. (2011, July). Particle swarm optimization for clustering semantic web services. In 2011 10th International Symposium on Parallel and Distributed Computing (pp. 170-177). IEEE.
- [38] Liang, Q., Li, P., Hung, P. C., & Wu, X. (2009, September). Clustering web services for automatic categorization. In Services Computing, 2009. SCC'09. IEEE International Conference on (pp. 380387).

IEEE.

- [39] Gholamzadeh, N., & Taghiyareh, F. (2010, December). Ontology based fuzzy web services clustering. In Telecommunications (IST), 2010 5th International Symposium on (pp. 721725). IEEE.
- [40] Gao, H., Stucky, W., & Liu, L. (2009, May). Web services classification based on intelligent clustering techniques. In Information Technology and Applications, 2009. IFITA'09. International Forum on (Vol. 3, pp. 242245). IEEE.