Similarity Analysis and Clustering for Web Services Discovery: A Review

Abdelmoniem Helmy
Department of Computer Science, Institute of Statistical Studies & Research (ISSR), Cairo University
Giza, Egypt.

Akram I. Salah
Department of C omputer Science, Faculty of Computers & Information Science(FCI), Cairo University
Giza, Egypt

Mervat H. Geith
Department of Computer Science, Institute of Statistical Studies & Research (ISSR), Cairo University
Giza, Egypt

ABSTRACT
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 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. In this paper the authors will provide a review for different similarity analysis approaches used for clustering web services into similar groups for benefit of service discovery.

General Terms
Web Services Discovery.

Keywords
Web Services, Services Similarity Analysis, Web Services Clustering, Service Discovery.

1. INTRODUCTION
When accessing the Internet, users want their search requests to be met with rapid and accurate service results. The growth in Internet popularity requires an accompanying increase in the efficiency and accuracy of the matchmaking process between user requests and web service discoveries. The process of retrieving relevant web services involves both the analysis and the classification of the available web services. Once services are analyzed and classified, they can be well-matched to user requests.

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 in order to accurately find services for a particular requested function. The main goal of clustering is to group objects in a way that objects in the same cluster have high similarity to one another, and at the same time are very different from objects in other groups.

Analyzing the functionality of Web services 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 [1]. Grouping services based on similar functionality could better match services with users' requests. This clustering approach would improve the quality of web service discovery and provide users with better-quality options in selecting a service.

There have been many applications of cluster analysis to practical problems. For example, in the context of Information Retrieval, Clustering can be used to group a Web search engine results into different categories where each category representing a particular aspect of the user's query [2]. The rest of the paper is organized as follow: In the next Section a taxonomy for web services similarity and clustering analysis from different angles is provided. In this context, three main similarity approaches for web services clustering are presented, namely structural similarity, lexical similarity and semantic similarity are presented in subsections 2.1, 2.2, and 2.3. Clustering and service recommendation is provided in section 3. Section 4 conclude this review.

2. APPROACHES FOR SIMILARITY ANALYSIS & CLUSTERING.
Figure 1. shows the taxonomy for web services analysis and clustering approaches. Two different kinds of clustering approaches have been proposed in past research for service clustering. One kind of approach is based on the information in non-semantic service descriptions and the other is based on semantic service descriptions. Non-semantic service descriptions can be used to cluster services into categories. However, the semantic approach can be used to compare new unclassified services with a set of already classified services. The approaches for clustering web services can be categorized into structural similarity clustering, lexical similarity clustering, and semantic similarity clustering. The next three sections will explain in detail these approaches.

2.1 Structural-Based Similarity Clustering
Mining WSDL documents content for clustering them into functionally similar web service groups is a key step in this process. Using the structure of the WSDL document, web service would be defined according to the service name, service description, the input set of service, and the output set of service. It would define the similarity of two services, and define the exact service request of a web service requestor. This approach can be categorized as follow.

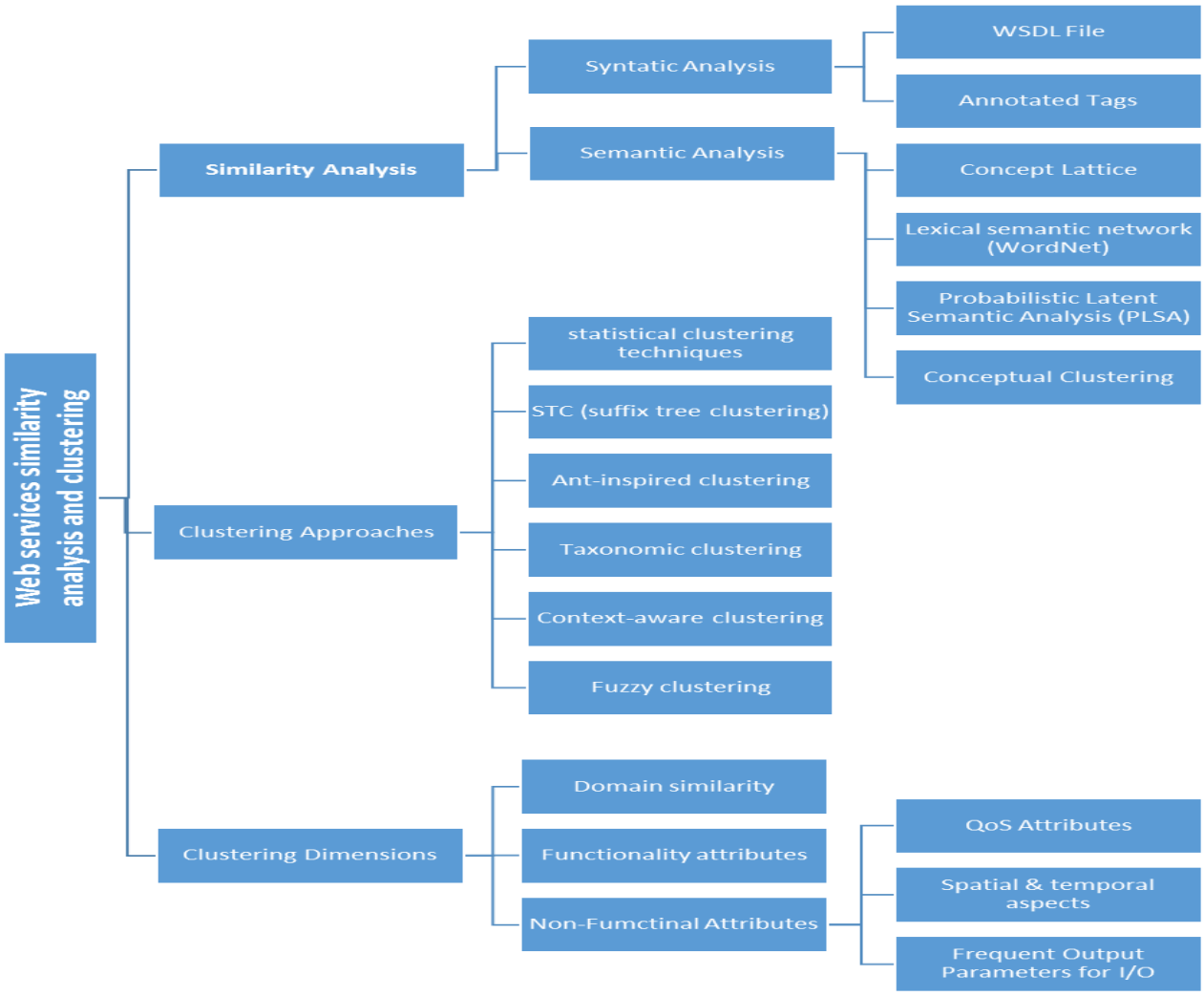

Fig 1: If necessary, the images can be extended both columns

2.1.1 Syntactic Structure Analysis

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 user requests [3]. 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. This process assists the search engine in identifying the functionality of the web service and better matching web services with users' requests. In this approach, five key features that can be extracted from WSDL documents and integrated in order to cluster web services into functionality-based groups are identified. These features are WSDL content, types, messages, ports (endpoints), and web service name. Integrating these features could create a more accurate representation of the functionality of a web service.

Approaches to web services clustering can also utilize structural similarity when analyzing the similarity of web services behavior. One case study illustrates the benefits of using process-based matchmaking of services and evaluates the strengths of different proposed heuristics [4]. The proposed approach uses internal process models of the services to achieve precision and recall performances, while providing accurate similarity degrees. In order to achieve this, the approach represents the underlying processes of web services and the requests for web services as finite state machines (FSMs). The matchmaking process combines structural similarity heuristics with a semantic similarity measure based on ontologies.

In order to determine the similarity of two services, structural heuristics compare the individual atomic processes of the services. The approach associates each atomic process with a semantic concept and compares the semantic similarity of processes. The matching results combine the structural similarity with semantic similarity in order to provide precise scores. If two atomic processes are identical, structural heuristics would assign 1 as the similarity value and 0 otherwise. This specification restricts the outcomes to exactly-matching processes. In order to discover processes that only partially match, the process relaxes the similarity of non-identical processes into a larger range of values. This approach requires that both services and requests are modeled as FSMs by developers and users. Developing an FSM of a service may be a complex action for a user; however, support tools can help users define their requests as FSMs in real-life applications of this approach.

2.1.2 Statistical Cluster Analysis

One research proposes a solution for web service discovery that combines semantic analysis with the techniques of

hierarchical clustering [5]. This approach would utilize the previously described process of cluster analysis of web services based on similarity. It would aim to improve the semantic web service discovery process based on a hierarchical clustering. The advantage of this approach is that it can produce a hierarchically-arranged clustering with a high level of accuracy. This process would first analyze web service description documents mathematically based on WSDL. Using the structure of the WSDL document, web service would be defined according to the service name, service description, the input set of service, and the output set of service. It would define the similarity of two services, and define the exact service request of a web service requestor.

Another research presents an approach that uses statistical cluster analysis [6]. This approach focuses on an efficient algorithm that can be scaled to large and distributed service repositories and can still guarantee an efficient processing of requests and subsequent clustering for generated matches. The usual distance for proximity measurement is extended by means of a multidimensional angle produced by a vector space search engine. Additionally, this research presents a more effective method for creating the matrix for the distance measurements. The proposed method is a modified version of common statistical cluster analysis. In this case, the cluster algorithm must not be limited to a specific number of variables or to a maximum size for the stored entries, in order to allow it to be executed on a multidimensional term space. The compatibility between the desired outcome and the expectations of a search request depends on the quality of the index as well as the algorithms used.

2.2 Lexical-Based Similarity Clustering

2.2.1 Suffix-Tree Clustering

Research has also evaluated approaches that use various other methods of calculating similarity. One research discusses a web service classification approach based on a suffix tree algorithm [7]. The proposed approach automatically classifies web services through semantic similarity calculation. The idea for service semantic classification consists of two modules: a document mining module and an automatic web service classification module. The first part document mining extracts important information from web service description documents. This involves two steps: deleting non-frequent words, and removing the pause and punctuation marks. The second part involves the service classification module.

The research adapts the traditional suffix tree clustering algorithm in order to solve several issues. In order to further merge base clusters to improve the accuracy of service classification, the proposed approach would utilize WordNet [8] to cluster synonymous nodes. Additionally, the approach can extract the clustering center to describe each cluster accurately. The classification module calculates the similarity of concept hierarchy in base clusters. The concept of similarity in the two base clusters is required to meet a certain threshold. This step excludes some non-similar base clusters. However, further analysis of instance hierarchy is required to assert whether two base clusters are truly similar or not. The module further calculates the similarity of keywords and phrases between two concepts in the instance hierarchy. If their similarity meets a certain threshold in the instance hierarchy, these two base clusters would be combined into a new base cluster. This approach would use a "concept and instance" tree structure and realize the similarity calculation via this structure. If web service is increasing exponentially, the suffix tree can increase linearly because of this tree structure. This feature would be able to deal with the rapidly growing web service of the Internet. The experimental evaluation conducted by this research indicated that this particular method of classification produces results efficiently.

2.2.2 Domain Classification

Another research expands on the use of lexical similarity in cluster analysis of web services. This research proposes a text mining approach to automatically classify services to specific domains and identify key concepts inside service textual descriptions [9]. This approach was validated in its implementation on a dataset of 600 web services yielding accuracy up to 90%. This classification approach can be used to focus user requests to a refined set of web service categories. The proposed approach uses WSDL features to classify web services. A model proposes that offline pre-processing classifies downloaded WSDL documents into functionally similar groups. Although a document could potentially be assigned to multiple categories, this approach assigns a single category to each document. Documents will be extracted for service name, service documentation, WSDL schema, WSDL messages, and WSDL port (endpoint) types.

2.2.3 WordNet-based Semantic Similarity

One research work [10] has evaluated the use of WordNet in analyzing the semantic similarity of web services. WordNet is an electronic lexical database that groups words into synonym sets and maintains semantic relationships between these sets. This database has proved instrumental in analyzing web service semantics by using word senses to determine underlying semantics. It is organized by meaning so that words in close proximity are semantically related. Typically, a single word can have different meanings based on the context in which it is used. Word sense disambiguation determines the correct sense of a word out of the multiple different meanings that could be potentially represented. In the study on the clustering of web services based on semantic similarity, researchers explored the semantics of web services using WSDL operation names and parameter names along with WordNet. WordNet is used to clearly determine a word sense in a particular context. An algorithm looks for all the paths from the context to the word and then selects the shortest path as the correct meaning. Six different measures are used to obtain the level of similarity between words in WordNet [11]. Once this approach has computed the semantic similarity of web services, it uses this data to generate clusters.

The work in [7] has expanded upon the use of semantic similarity evaluation through WordNet. The approach used in this research applied different algorithms to the semantic similarity calculation. This demonstrates how WordNet can be utilized in the application of various techniques for semantic similarity calculation. This particular use of WordNet allowed for the development of an approach that automatically classifies web services. The approach proposed a "concept and instance" tree structure in order to achieve the similarity calculations. The rapid growth of web services requires that the classification process adapt to processing larger amounts of web services. The application of WordNet can contribute to a more efficient classification process.

2.3 Semantic-Based Similarity Clustering

Research on web service discovery has demonstrated the advantages of utilizing the semantic meanings of web services in the matchmaking process. The classification and discovery process demonstrates that in order to add semantic meaning to web services, researchers must first classify web services into different categories. Classifying web services efficiently and accurately prevents mistakes in matching services to requests.

This section discusses three techniques to make use of semantic meaning of web service content.

2.3.1 Adapted Ontology Service Capability Similarity

Currently, the semantic languages for services, such as the OWL-S and the WSMO, are required for semantically representing service capabilities. Therefore, service discovery focuses on the matchmaking of service capability rather than ontology-based service selection. One experiment aims to remedy this perceived shortcoming by combining two adapted methods in order to calculate the semantic distance of single formal ontology concepts [12]. The two approaches are a fuzzy-weighted associative network (an edge-based measure) and an information-theoretical approach (content-based measure). This research utilizes previous measures used for ontology similarity in order to develop a new and more efficient service discovery method. This combined approach aims to determine the ontology similarity of various semantic services. First, it differentiates two cases of service ontology concepts single and compound ontology concepts in order to measure their similarity in service context. Then, the approach would implement an adapted ontology-similarity measurement that combines three existing methods.

The research on this approach defines a new model for service discovery based on ontology similarity. This model improves service selection outcomes. Research has concluded that an ontology similarity-based approach to the measurement of service similarity works well for the experimental data. The research analyzes the ontology similarity problem in a semantic service context, classifies the ontology concept name features used by service description, and presents an adapted ontology concept distance method in order to further the measure of service similarity.

2.3.2 Combined Function and Text Similarity

Another mode of web services analysis primarily involves semantic analysis. One approach in particular utilizes certain techniques that solve deficiencies demonstrated by previous works [13]. First, the approach combines TF/IDF, an information-retrieval method, and ontology, or semantics, to effectively cluster web services; second, it uses Ward's distance to expedite recommendation; and third, it builds a service space model that serves to propose a service recommendation method. When cluster analysis of web services, this approach evaluates similarity based on both functionality similarity and text similarity. To discover the similarity of text, the approach uses TF/IDF technology to weigh the importance of terms in the body of text. TF/IDF will determine the terms that appear more frequently in a specific text description of a web service than other in other descriptions.

This approach then utilizes the previously described method of hierarchical clustering. The key operation of hierarchical clustering is the computation of the distance between two clusters. Different definitions of the distance between clusters lead to different algorithms. The approach uses Ward's distance rather than other measures because it is less susceptible to noise and outliers. Finally, this approach proposes a service space model to be implemented as part of the service recommendation method. This method recommends component services with similar topics and good performance to composite services. It uses matrix decomposition, which is able to recover incomplete information, to recommend component services to composite services. This method of service clustering and recommendation is proposed for the convenience of service usage and service composition. The recommendation model, or service space model, could be extended in multiple dimensions.

2.3.3 Semantic Similarity with Hierarchical Clustering

One research specifically investigates the clustering of web services based on semantic similarity [14]. The approach utilizes the novel idea of representing clusters by the most similar operations of web services in that cluster. The research develops an application for effectively finding similar or related web services, which can be used as an add-on to any web service search engine with UDDI repository. The research used semantics of WSDL along with WordNet to compute similarity between web services. For web services in the sample set of test data, hierarchical clustering was used to compute similarities and cluster data. Each cluster was represented by a set of characteristic operations. These cluster representations were used to evaluate similarity of any new web services using the nearest neighbor approach. In this method, each web service is originally treated as belonging to a cluster. Then, a similarity matrix of web services determines the nearest neighbors. Nearest clusters are then merged into one cluster. This process is repeated until all the web services merge to a single cluster. This process yielded good results and an accuracy of 70% in the case of the test data.

3. SERVICE CLUSTERING FOR SERVICE RECOMMENDATION

The accurate prediction of similar web services can greatly boost the recommendation process. Research on the clustering and recommendation for semantic web services in time series demonstrates how recommendation technology can select possible component services to composite services based on the history information of a user's invocations and similar services. Therefore, the accurate prediction of similar web services can successfully determine which services a user would likely want to use based on past use. Predicting service use relies on an accurate evaluation of the similarity between services, as demonstrated earlier.

The development of newer versions of old services can actually deter the search discovery process. Some web services become outdated or even obsolete due to the production of new versions, and some services work well only with other services of older versions. These improper services lower the performance of the whole service they belong to. Research has proposed an approach that remedies these problems [15]. This approach includes both a clustering method and a recommendation method. Clustering technology classifies semantic services according to their topics and functionality. In contrast, recommendation technology predicts the possible preference of a composite service and recommends possible component services to the composite service according to the history data of similar composite services. Experiments demonstrate that this particular clustering method, combined with ontology and TF/IDF technology, produces more accurate results than other approaches. They also demonstrate that this recommendation method produces less average error than other approaches.

4. CONCLUSIONS

The growth in Internet popularity requires an accompanying increase in the efficiency and accuracy of the matchmaking process between user requests and web service discoveries. In order to make these improvements, researchers must

formulate and test new approaches to the service discovery process. Analysis and clustering of web services based on functional similarities would greatly improve the ability of search engines to retrieve the most relevant web services in response to user requests.

Although, the prospects of discovering services have considerably improved with the appearance of semantic Web services, the existence of a large number of available services makes the discovery process inefficient, unless services are grouped according to specific criteria [16]. Research in cluster analysis of web services has recently gained much attention 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.

The above researches studied the hybrid solution for building and filling a taxonomy structure of web services by combining semantic analysis with clustering techniques. It concluded that the method of web services discovery based on clustering could be devised to improve the efficiency and accuracy of web service discovery. The retrieval process would benefit from improved clustering techniques. It can significantly reduce the overhead and enhance the service discovery efficiency However, a full understanding of web service potentiality can only be achieved when the effects of contextual interpretation can be made by the service provider and consumer through pragmatic consensus.

5. REFERENCES

[1] Cong, Z., Fernandez, A., Billhardt, H., & Lujak, M. (2015). Service discovery acceleration with hierarchical clustering. Information Systems Frontiers, 17(4), 799-808.

[2] Poblete, B. (2010). Query-based Data Mining for the Web. Universitat Pompeu Fabra.

[3] Elgazzar, K., Hassan, A. E. & Martin, P. (2010). "Clustering WSDL to Bootstrap Discovery of Web Services." International Conference on Web Services, 147-54. Print.

[4] Gunay, A., & Yolum, P. (2007). "Structural and Semantic Similarity Metrics for Web Service Matchmaking." E-Commerce and Web Technologies 4655 (2007), 129-38. Print.

[5] 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. 242-245). IEEE.

[6] Platzer, C., Rosenberg, F., & Dustdar, S. (2009). Web service clustering using multidimensional angles as proximity measures. ACM Transactions on Internet Technology (TOIT), 9(3), 11.

[7] Deng, F. (2012). "Thesis: Web Service Matching based on Semantic Classification." School of Health and Society, Department of Computer Science.

[8] Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11), 39-41.

[9] Nisa, R., & Qamar, U. (2014). "A text mining based approach for web service classification." Information Systems and e-Business Management. Print.

[10] Konduri, A. (2008). Clustering of Web Services Based on Semantic Similarity (Doctoral dissertation, University of Akron).

[11] Meng, L., Huang, R., & Gu, J. (2013). A review of semantic similarity measures in wordnet. International Journal of Hybrid Information Technology, 6(1), 1-12.

[12] Wang, X., Ding, Y. & Zhao, Y. (2006). "Similarity Measurement about Ontology-based Semantic Web Services." In Proceedings of Workshop on Semantics for Web Services. Print.

[13] Lei, Y., Wang, Z., Meng, L., & Qiu, X. (2014). Clustering and Recommendation for Semantic Web Service in Time Series. TIIS, 8(8), 2743-7362.

[14] Cong, Z., & Gil, A. F. (2013). "Efficient Web Service Discovery Using Hierarchical Clustering." Agreement Technologies 8068, 63-74. Print.

[15] Deng, S., Wu, Z., Wu, J., Li, Y., & Yin, J. (2009). An efficient service discovery method and its application. International Journal of Web Services Research (IJWSR), 6(4), 94-117.

[16] Chifu, V. R., Pop, C. B., Salomie, I., Dinsoreanu, M., David, T., & Acretoaie, V. (2011). Ant-based Methods for Semantic Web Service Discovery and Composition. Ubiquitous Computing and Communication Journal, 631-641.