Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/130107

An Analysis on Web-Service-Generated Data to Facilitate Service Retrieval

I.R. Praveen Joe[1,*] *and P. Varalakshmi*[2]

[1] Department of CSE, KCG College of Technology, Chennai, India
[2] Department of CT, MIT, Anna University, Chennai, India

Received: 8 Jun. 2018, Revised: 20 Aug. 2018, Accepted: 25 Aug. 2018
Published online: 1 Jan. 2019

Abstract: In this paper, a novel layered clustering approach is proposed to cluster web services in order to facilitate web service selection process. The process of web service selection from a rapidly growing number of functionally similar services in the internet, results in an increase of service discovery cost, transforming data time between services and service searching time. Though suitable technologies for web services clustering are being developed, blending neural networks and swarm-based algorithms is not prevailing. A novel two-phase clustering approach involving ART (Adaptive Resonance Theory) network for primary clustering with functional data and swarm algorithms (BOIDs, ABC and PSO) for sub-clustering with non-functional data (metadata, QoS and service-generated data respectively) is proposed in this work. As a result of this layered approach, the computational overhead is greatly reduced and the search space is also abridged significantly in order to obtain optimal services.

Keywords: Web Service Selection, Web Services Clustering, ART network, Swarm algorithms.

1 Introduction

Numerous disorganized web services are existing in the internet in the form of service repositories and it is very tough for the user to choose the required web service of his or her priority. This raises the service discovery cost, transforming data time between services and service searching time [1]. While automating the service selection process, we need to carefully involve both functional and non-functional requirements which are explicitly stated and implicitly understood. Functional requirements mainly focus on functionality, given a booking service; an example of functional requirements is that a flight ticket with price less than $ 1000 is preferred. Non-functional requirements are not explicitly specified, in the example of the booking service, non-functional requirement is that the service should respond to the user within the response time of 5 seconds.

Though keyword search techniques are popular, these methods are confined to a search within the service description of the web service and it is practically impossible to accommodate fine details of the requirement in a few words. Therefore, for creating an automated service selection recommendation system, it is essential to categorize the services in an appropriate way that results in reducing the search space and time.

While making such an attempt to operate on an efficient clustering technique, the efficiency of clustering can be attributed through four main aspects. They are (a) services representation (in this research three categories of data namely meta data, QoS data and service-generated data are used) (b) similarity measure (in this research compactness (intra-cluster distance), isolation (inter-cluster distance) and Dunn Index are used) (c) choice of algorithms (in this research a novel and hybrid approach employing ANN and three swarm algorithms is used) (d) Reducing the search space (in this research a two-phase filtering approach is used where sub clustering is applied over the shortlisted set of services with minimum homogeneity).

In this paper, experiments that have been made with service-generated data are elaborated. This refers to trace logs, domain logs, service relationships etc. These data are recorded over a period of time after the service became completely functional. There are two main reasons for employing service-generated data. The first one is the availability of all meta data documents with all relevant filled up elements may not be guaranteed. And

* Corresponding author e-mail: praveenjoeir@outlook.com

the second one is, QoS data for the same service vary from machine to machine as they are influenced by network conditions. Therefore, data like domain logs, service relationships, etc., recorded for a definite-time period are observed to be more reliable and stable.

2 Related Work

A literature survey has been conducted on the various approaches available for web services clustering. In such a study, it is generally understood that clustering can be separated into two sub groups. Hard Clustering: in hard clustering, each data point either fit in to a cluster absolutely or not [2]. Soft Clustering: in soft clustering, instead of positioning each data point into a separate cluster, a probability or likelihood of that data point to be in those clusters is assigned [3].

Also clustering algorithms are compared by means of two-cluster validation measures: internal measures, which exploit intrinsic information in the data to evaluate the quality of the clustering. Internal measures include the connectivity, the silhouette coefficient and the Dunn index [4]. Stability measures a special adaptation of internal measures, which appraise the consistency of a clustering outcome by comparing it to the clusters obtained after each column is removed, one at a time [5]. Some of the familiar approaches are studied as below.

Domain ontology-based approaches for web service clustering attempts to lessen the manual discovery and usage of web service by permitting software agents to spontaneously and dynamically retrieve web services. [6] It is an operative and consistent technique. Relevancy of web service discovery can be amended by augmenting semantics by means of expressive formats like OWL [7]. They entail the end user to have a wide knowledge of semantic web services, related description and implementation details are problematic for the end users [8]. The discovery opportunity of these methods is frequently restricted to some web services that are published in a specific description standard. The service requestor may not be conscious of all the footings related to the service request. Ontology mapping practices may be used to coordinate the differences between these ontologies for backing interoperability [9].

Public ontology-based approaches permit developers to improve web services with semantic information without semantic annotation against ontology [10]. WordNet is not domain specific and disregards the semantic annotation cost of services. Terms and concepts in WordNet has its detailed semantic. Different parts of WordNet have dissimilar granularity for the description of word senses. In general, WordNet is too fine-granular for many situations. There is no real multi-lingual WordNet. Syntax and semantic-based approaches are well suited and are extensively used. Standards like UDDI exist. Keyword-based search is more acquainted to the user but requires human interaction and can't select most parallel service among a large set of available and semantically similar services [11].

Context-aware approaches do empower the involuntary discovery of distributed web services grounded on comprehensive semantic representations. Context-awareness is a key ingredient in any ubiquitous and pervasive system and delivers intelligence to the system. The concept of context is quite expansive. Due to the intricacy of the context, it is infeasible to appropriately model context for all applications [12].

Swarm algorithms are observed to be suitable in this context of the research because it is applied in the second phase for sub-clustering in all the three stages for the reason that there exists a homogeneity in the clusters formed in phase one and all swarm intelligent algorithms are applied on similar sort of agents. For example, there are no two different kinds of fishes school, no two species of birds flock together and so on. The other factors [13] that help us to use swarm algorithms over other algorithms are

flexible: the colonies respond to internal disturbances and external challenges;
robustness exists: tasks are completed even if some agents fail;
scalability exists: they are scalable from a few agents to millions;
decentralized control: there is no central control in the colony;
self-organized: the solutions are emergent rather than pre-defined;
adaptation possible: the swarm system cannot only adjust to predetermined stimuli but also to new stimuli;
higher speed possible: changes in the network can be propagated very fast;
modularity exists: agents act independently of other network layers;
parallelism exists: agents' operations are inherently parallel.

BOIDS, ABC and PSO work hand-in-hand with ART algorithm. The following facts help us to substantiate how well ART- and swarm-based algorithms complement each other in the context of web services clustering. Firstly, Non-linearity is handled in all the algorithms [14]. They withstand high dimensional data values; also they evolve weight calculations and use fitness functions in common [15]. Computational overhead is controlled due to the two-phase approach. For example, optimality is achieved in a better way in PSO as there is no overlapping and mutation.

When ART network does not address optimality, swarm algorithms help in achieving optimal results apart from clustering [16]. Swarm algorithms do not have any vigilance value or threshold to control clustering and for this reason [17], ART algorithm compensates this by taking observations for varied vigilance values before sub-clustering [18]. This is ensured with the percentage of relevancy in each stage. No two different kinds of agents work together [19]. So, swarm algorithms are suitably

applied upon achieving minimum homogeneity, after clustering through ART in the first phase of all the three stages of research. All the three swarm algorithms use competitive learning strategy and ART algorithm uses incremental learning approach. Thus, it is claimed that the approaches are hybrid in nature and there exists a novelty.

Table 1: Comparison of ART with other clustering methods

Methods	No. of Clusters	Ave. Intracluster Distance	Ave. Intercluster Distance
Hierarchical	7	23.412 to 69.711	21.223 to 49.556
Decision Tree	7	21.451 to 62.454	28.531 to 51.454
Naï ve Bays	8	19.456 to 56.734	35.334 to 57.841
K Means	10	18.565 to 49.456	38.532 to 59.342
ART	12	15.432 to 34.165	41.643 to 60.421

3 Proposed Work

The main objective of this research work is to focus on service discovery which means selecting services that are more required based on their relevancy from an already prepared set of services. One of the data mining techniques namely clustering is used in order to facilitate the selection of relevant web services. Clustering of services is done through effective clustering algorithms and suitable web services are elected based on their relevancy. It considers not only user's expectations (or) needs but also all the possible non-functional information, so that the service retrieval process is executed in a more effective manner.

In the presented experiment ART (Adaptive Resonance Theory) is used for primary clustering and PSO (Particle Swarm Optimization) is used for sub clustering. ART is an unsupervised algorithm that tackles with the stability and plasticity problems. This means similar to that of the human brain, the network learns on an incremental fashion and at the same time it could retain the old information as and when accommodating new information. ART algorithm requires setting up of a threshold value namely vigilance parameter ρ which decides the degree of similarity. If $\rho = 0$, then no clusters are formed and if $\rho = 1$, then two data values need 100% match to fall in a cluster.

In an earlier approach, the credibility of ART algorithm is alone tested with an input set of 951 web services considering only functional requirements and ART proves to give more passable results than others as shown in Table 1. ART networks consist of an input layer and an output layer. Bottom-up weights are used to determine output-layer candidates that may best match the current input. Top-down weights represent the "prototype" for the cluster defined by each output neuron. A close match between input and prototype is necessary for categorizing the input. Finding this match requires multiple signal exchanges between the two layers in both directions until "resonance" is established or a new neuron is added.

In the proposed PSO algorithm with MapReduce, the clustering task is formulated as an optimization problem to obtain the best solution based on the minimum distances between the data points and the cluster centroids. This is a partitioning clustering algorithm similar to the k-means clustering approach, in which a cluster is represented by its centroid. In k-means clustering, the centroid is calculated by the weighted average of the points within a cluster. But here, the centroid for each cluster is updated based on the swarm particles' velocities. Moreover, each particle Pi contains information which is used in the clustering process such as: Centroids Vector (CV): current cluster centroid's vector.

- Velocities Vector (VV): current velocities vector.
- Fitness Value (FV): current fitness value for the particle at iteration t.
- Best Personal Centroids (BPC): Best personal centroids seen so far for P_i
- Best Personal Fitness Value (BPC_{FV}): Best personal fitness value seen so far for P_i
- Best Global Centroids (BGC): Best global centroids seen so far, for whole swarm.
- Best Global Fitness Value (BGC_{FV}): Best global fitness value seen so far, for whole swarm.

This information is updated in an iteration-based on the previous swarm state. The two main operations need to be adapted and implemented to apply the clustering task on a large-scale data: the fitness evaluation, and the particle centroids updating. Particle centroid updating is based on PSO movement given in Eqs. (3.1) and (3.2) that calculates the new centroids in all iterations for the individual particles. Particles are shifted in the problem search space as given in Eq. (3.1).

$$X_i(t+1) = X_i(t) + V_i(t+1) \tag{3.1}$$

where X_i is the position of particle i, t is the iteration number and V_i is the velocity of particle i.

Particle velocities are updated in PSO algorithm as given in Equation (3.2).

$$\begin{aligned} V_i(t+1) = W.V_i(t) + (r_1.cons_1).[XP_i - X_i(t)] \\ + (r_2.cons_2).[XG - X_i(t)] \end{aligned} \tag{3.2}$$

where W is inertia weight, r_1 and r_2 are randomly-generated numbers, $cons_1$, $cons_2$ are constant coefficients, XP_i is the current best position of particle i and XG is the current best global position for the whole swarm.

Besides the update of the particle centroids, the fitness evaluations are based on a fitness function that measures the distance between all data points and particle centroids by taking the average distance between the particle centroids. The fitness value is evaluated as given in

Eq. 3.3.

$$Fitness = \frac{\sum_{j=1}^{k} \frac{\sum_{i=1}^{n_j} Distance(R_i, C_j)}{n_j}}{k} \quad (3.3)$$

where n_j denotes the number of records that belong to cluster j; R_i is the ith record; k is the number of available clusters; $Distance(R_i, C_j)$ is the distance between record R_i and the cluster centroid C_j. Later Manhattan distance is calculated as given in the Eq. 3.4.

$$Distance(R_i, C_j) = \sum_{\upsilon=1}^{D} \left| R_{i\upsilon} - C_{j\upsilon} \right| \quad (3.4)$$

where D is the dimension of record R_i; $R_{i\upsilon}$ is the value of dimension υ in record R_i; $C_{j\upsilon}$ is the value of dimension υ in centroid C_j.

Thus, PSO is a population-based search technique where the individuals are denoted as particles and are clustered into a swarm. Each particle in the swarm denotes a candidate solution to the optimization problem which is flown through the multidimensional search space. The particles then modify their position in search space allowing to their own experience and that of neighbouring particles. A particle therefore makes use of the best position met by it and the best position of its neighbours to position itself near the best solution. This type of bio-inspired clustering provides compact and accurate clusters. Here, PSO-based clustering is done in parallel. Parallelism can be attained by using Map Reduce technique in Hadoop environment. Map Task is used to find pbest and gbest values and the Reduce Task is used to find the gbest value out of the outcomes of gbests from mappers. Parallel PSO clustering provides faster processing. From the clusters formed, relevant observations are made and reports are extracted by refinement. The report is presented in such a way that, it is easy for the user to understand and choose a suitable service. The architecture of the proposed model is shown in Fig. 1.

4 Results and Discussions

The data involved in the study is broadly classified as functional and non-functional data. Every feature set is a combination of bits which characterize a web service.

4.1 Definition of Functional Parameters

Web Service (ws): ws = {Name, Input, Output, Operation} where Name refers to the name of ws, Input = $\{IN_i, IN_i\} \in \{\text{Entity Set}\}$, $i \in \{1,2,\ldots,n\}$, Input refers to the input characteristics of the web service, Output = $\{OUT_j, OUT_j\} \in \{\text{Entity Set}\}$, $j \in \{1,2\ldots,m\}$, Output refers to the output characteristics of web service and Operation = $\{OP_k, OP_k\} \in$ Entity, $k \in \{1,2,\ldots,l\}$, Operation refers to the operational characteristics of the web service

Table 2: Data fields of input for feature construction

Field. No	Field Name	Data
1	domain_name	okjobmatch.com
2	query_time	14.7
3	created_date	1/2/2017
4	update_date	1/2/2017
5	expiry_date	1/2/2019
6	domain _registration_id	146
7	domain_register_name	godaddy.com
8	domain_register_who_is	whois.godaddy.com
...	...	...
50	name_server1	ns1.dynadot.com
51	name_server2	ns9.rookdns.com
52	name_server3	ns7.dynadot.com
53	name_server4	ns8.rookdns.com
54	domain_status1	client delete prohibited
55	domain_status2	client renew prohibited
56	domain_status3	client transfer prohibited
57	domain_status4	client update prohibited
58	domain_status _code	200

4.2 Non-Functional Parameters

The first set of data are elements of Meta data documents like XML Schema, XSDL, WS-Policy, WS-Addressing, WS-Meta data Exchange. The second set of data are QoS parameters like response time (negative QoS, i.e., decreasing values are preferred) and throughput (positive QoS, i.e., increasing values are preferred) and the third set of data are service-generated data like domain logs, trace logs, service relationships

In this research work, the services that offer domain name purchase for hosting their websites have been considered. The source of the dataset is www.databasereseller.com. Around two lakh records are considered for the experiment. As discussed earlier, in all the three stages of the research work, functional data and non-functional data have been employed (varied at every stage). A snapshot of the data pertaining to purchase of domain names is shown in Table 2. There are a total of 58 fields i.e. dimensions for every data item. This is helpful in quantifying the input feature set.

These values are normalized appropriately and given to ART algorithm which accepts binary data only. An example of quantification of data into binary values is shown in Table 3.

Likewise the bit patterns may be generated and every feature matrix as per the need could be constructed. The inputs are provided to the clustering algorithm ART and later sub-clustering is done through swarm optimization techniques. Let X and Y be the bit patterns generated for 2 web services which are in the same cluster and the Jaccard

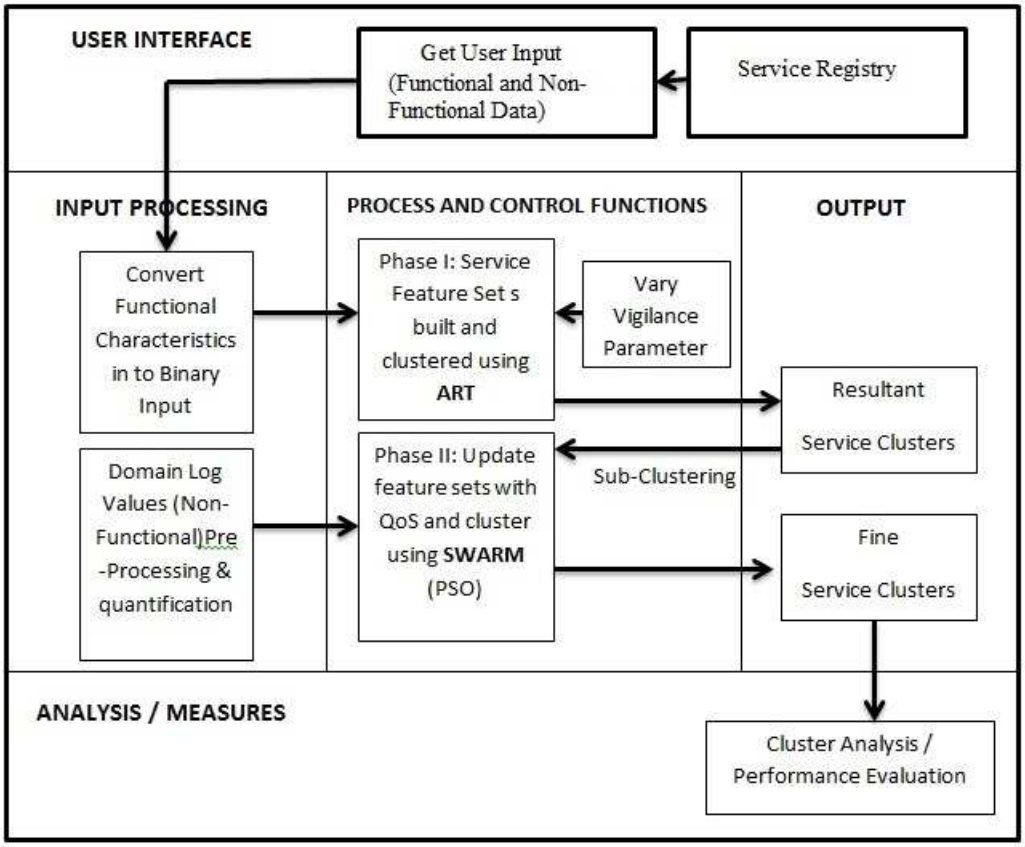

Fig. 1: Architecture of the system model

Table 3: Quantification of input data in to bit patterns

Field	Feature Bit
Domain Status Code	Equal to 200 means 1(Preferable) Equal to 404 means 0 (Not Preferable)
Query Time	Below 15 means 1 (Preferable) Above 15 ms means 0 (Not Preferable)
Created Date	After 1.1.2017 means 1 (Preferable) Before 1.1.2017 means 0 (Not Preferable)
Expiry Date	More than one year means 1 (Preferable) Less than one year means 0 (Not Preferable)
Domain label name	Less than 63 characters means 1 (Preferable) More than 63 characters means 0 (Not Preferable)
Billing Country	Equal to India means 1 Any other country means 0

distance is computed as below:

$$X = \{1,0,1,0,0,0,1,1,1,1,0,1\}$$
$$Y = \{1,0,0,0,1,1,1,0,1,1,0,1\}$$

M_{00} = No. of attributes with X and Y both 0
M_{01} = No. of attributes with X as 0 and Y as 1
M_{10} = No. of attributes with X as 1 and Y as 0
M_{11} = No. of attributes with X as 1 and Y as 0

Jaccard coefficients $JC = (M_{11})/(M_{01} + M_{10} + M_{11}) =$ number of M_{11} matches / number of not-both-zero attributes

$$JC = (5)/(2+2+5) = 5/9 = 0.556.$$

Similar values are computed between X and every other value of Y in the cluster. Their average is taken as the average of intra-cluster distance. So, every cluster then has one minimum value of intra-cluster distance and one maximum value of intra-cluster distance. Similarly, the Jaccard distance is calculated between the centroids of two clusters to get the inter-cluster distance. This is implemented with the help of MATLAB software.

Table 4 shows the observed intra- and inter-cluster distances while clustering through ART with PSO, keeping the input set fixed (951 records) and varying the vigilance parameters (ranging from 0.5 to 0.8).

Table 4: Sample Iterations—for fixed input and varied vigilance values using ART and PSO

Iteration	v.p.	No. of inputs	No. of clusters	Ave. interacluster distance	Ave. intercluster distance
1	0.8	951	23	7.445 to 24.343	50.123 to 72.432
2	0.7	–	24	11.234 to 23.212	45.435 to 71.323
3	0.6	–	20	13.455 to 26.432	42.323 to 67.432
4	0.5	–	17	19.454 to 30.543	41.343 to 65.457

Table 5: Sample Iterations—Comparison of results for varied input and fixed vigilance value using ART and PSO

Iteration	v.p.	No. of inputs	No. of clusters	Ave. interacluster distance	Ave. intercluster distance
1	0.8	951	22	7.342 to 27.654	50.176 to 73.234
2	–	1903	25	11.245 to 27.564	52.543 to 74.321
3	–	2854	29	17.453 to 30.234	54.244 to 75.345
4	–	3807	35	22.345 to 34.213	56.356 to 76.345

Table 6: Comparisons of ART and PSO approach with other two approaches

	ART with PSO (Domain Logs)			ART with ABC (QoS)			ART with BOIDS (Meta Data)		
No. of Services	No. of Clusters	Ave. Intra-cluster Distance	Ave. Inter-cluster Distance	No. of Clusters	Ave. Intra-cluster Distance	Ave. Inter-cluster Distance	No. of Clusters	Ave. Intra-cluster Distance	Ave. Inter-cluster Distance
951	22	07.342 to 27.654	50.176 to 73.234	21	08.122 to 28.432	49.543 to 71.345	18	09.911 to 30.251	48.618 to 69.451
1903	25	11.245 to 27.564	52.543 to 74.321	23	12.275 to 30.521	50.114 to 66.432	20	14.145 to 32.043	49.564 to 65.513
2854	29	17.453 to 30.234	54.244 to 75.345	27	19.453 to 33.176	52.312 to 69.123	25	17.535 to 34.123	51.873 to 68.136
3807	35	22.345 to 34.213	56.356 to 76.345	33	26.734 to 35.353	55.123 to 71.234	31	24.812 to 37.541	53.618 to 69.451

Table 5 shows how unsupervised algorithms produce passable results when the number of input services is increased for a fixed vigilance value of 0.8.

The results obtained and represented in terms of intra- and inter-cluster distances for ART with PSO is compared to the previous two approaches namely ART with ABC and ART with BOIDs and are shown in Table 6.

Performance analysis through an evaluation index called Dunn index is explained below.

The Dunn index is used to recognize dense and well-separated clusters. It is defined as the ratio between the minimal inter-cluster distances to maximal intra-cluster distance. For each cluster partition, the Dunn index can be calculated as shown in Eq. 4.1.

$$D = \frac{\min\ 1 \le i \le j \le i\ d(i,j)}{\max\ 1 \le k \le n\ d'(k)} \tag{4.1}$$

where $d(i,j)$ represents the distance between clusters i and j, and $d'(k)$ measures the intra-cluster distance of cluster k. The inter-cluster distance $d(i,j)$ between two clusters may be any number of distance measures, such as the distance between the centroids of the clusters. Similarly, the intra-cluster distance $d'(k)$ may be measured in a variety of ways, such as the maximal distance between any pair of elements in cluster k. Since internal criteria seek clusters with high intra-cluster similarity and low inter-cluster similarity, algorithms that produce clusters with high dunn index are more desirable.

Table 7 shows the dunn index values for various clusters in iteration 1. It also depicts that, the cluster 9 contains largest dunn index value which is 0.956. So, cluster 9 is a well separated cluster when compared to the others. In the above instance, 9 out of 16 clusters have higher values.

When the entire data set is subjected to experiment, stage three produced clusters with the higher values of dunn index. 75% of the clusters have a dunn index of 0.7 and above, where as in the stages 1 and 2, they are 62.5% and 68.75% respectively.

Table 8 depicts 5 different clusters each containing homogenic web services. The services are grouped based on specific characteristics. For instance, cluster 1 which contains 7 services that are related to vehicles. Cluster 2 contains 7 services that are related to entertainment-based content. Cluster 3 contains services related to social

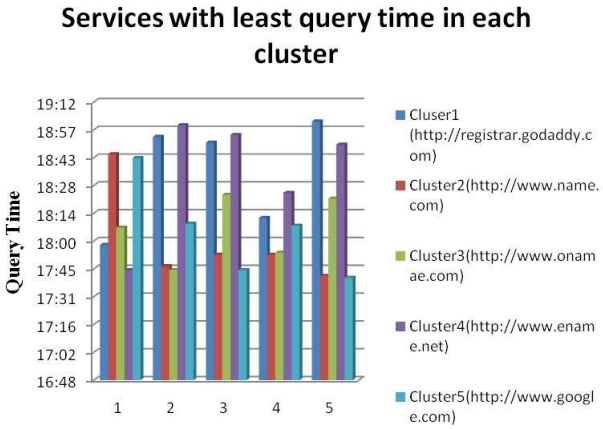

Fig. 2: Graph depicting web services with least query time

awareness, social discipline, and social activities like scout and NSS camp. Cluster 4 contains 5 similar services related to 'Miss Teen' awards of different cities and Cluster 5 contains blog related contents.

$$\text{Credit} = \frac{1}{(\text{in-degree} \times \text{number of activities})}$$

$$\text{(or)} \qquad = \sum_{0}^{n} \frac{1}{(\text{in-degree}_n \times \text{number of activities}_n)} \tag{4.2}$$

if out-degree is varying from 0 to n.

Table 7: List of Dunn index values for the three schemes

Cluster Number	Dunn Index		
	Stage 1	Stage 2	Stage 3
1	0.723	0.735	0.834
2	0.667	0.689	0.712
3	0.679	0.785	0.845
4	0.364	0.467	0.432
5	0.452	0.479	0.523
6	0.707	0.715	0.757
7	0.701	0.715	0.765
8	0.012	0.014	0.018
9	0.713	0.739	0.956
10	0.779	0.781	0.856
11	0.112	0.114	0.115
12	0.716	0.719	0.818
13	0.792	0.799	0.934
14	0.735	0.742	0.745
15	0.711	0.719	0.732
16	0.799	0.812	0.854

From every potential cluster, services can be choosen based on the priority.

Fig. 2 shows a graph that highlights the optimal services based on query time from each of the clusters namely cluster1 to cluster 5 that are shown in Table 8. For example in cluster 1, godaddy.com is the optimal service meant for domain name purchase service in India amidst

Table 8: Sample of cluster

Cluser1	20westband.com 2100southocean.com 2017filmler.com 210autohaus.com 210exoticrentals.com 210exotics.com 210liftedtrucks.com
*	21189sailorsbay.com
Cluster2	24losangelesolympics.net 208world.com 20pho7daily.com 20percent20th.org 20percent20th.com 20pearlsgirls.com
Cluster3	20mg5mgcialis.com 20gamlps.org 2017hyundaisonata.com 2017ldsencampment.com 2017ldsencampment.net 2017missteenashville.com 2017missteencharleston.com 2017arichtech.com
Cluster4	2017missteencharlotte.com 2017missteencolumbia.com 2017missteenfayetteville.com 2017missteengreensboro.com 2017missteengreenville.com
Cluster5	2017missteenraleigh.com 2020takeback.com 2020ryan.com 2020nhprimary.com 2020election.org 2018carprice.com

Table 9: Credit ranks of shortlisted services

Services	Computed Credit	Credit Rank
1	0.45677	13
2	0.46789	11
3	0.78271	8
4	0.45698	12
5	0.97673	3
6	1.00000	1
7	0.97589	4
8	0.10923	15
9	0.67875	10
10	0.78654	7
11	0.97332	5
12	0.98753	2
13	0.76278	9
14	0.86423	6
15	0.13456	14
16	0.10611	16

all other services in the cluster while considering query time which is 17.42 ms. Services are ranked based on the service credits. In a directed network, we have in-degree

and out-degree. In-degree is the amount of links directed to the node and out-degree is the amount of links that the node directs to others. Thus, credit can be measured using in-degree, out-degree and a number of activities after constructing the service network model using the formulae in Eq. 4.2 and the ranking is shown in Table 9.

5 Conclusion

In this paper, users are given recommendations about the web services available for the specific application by analyzing the service-generated data such as domain logs of web services by performing parallel PSO-based clustering using Map Reduce technique and then extracting reports by further refinement to achieve optimality. Application of ART algorithm has supported significantly in handling different classes of unfragmented data and it has been made possible to incorporate multidimensional non-linear data as ART algorithm-accepted binary inputs. Medium range of vigilance values between 0.5 and 0.8 is used to set the threshold values for the degree of similarity and it is observed that the cluster parameters are phenomenally increasing both in number and quality.

Usage of PSO algorithm has helped to achieve optimality in a faster rate and the computational overhead is controlled significantly due to the two-phase approach. Optimality is achieved in a better way in PSO as there is no overlapping and mutation. Formation of outliers was also avoided. The self-organizing and scalable nature of swarm algorithms complement unsupervised clustering of ART.

As a future work, it is planned to consider other service-generated big data such as service relationships, and recommendation reports could be obtained and compared to the results obtained by analyzing domain logs. Also convolutional neural networks may be used to extract data in the future.

References

[1] S.K. Rangarajan, V. Phoha, K. Balagani, R.R. Selmic and S.S. Iyengar, Web user clustering and its application to prefetching using ART neural networks, Journal of Computers, Vol. 15, No. 5, pp. 45–62 (2004).
[2] K. Su, B. Xiao, B. Liu, H. Zhang and Z. Zhang, TAP: A personalized trust- aware QoS prediction approach for web service recommendation, Journal of Knowledge-Based Systems, Vol. 115, No. 9, pp. 55–65 (2017).
[3] W. Wong, Liu and M. Bennamou, Tree traversing ant algorithm for term clustering based on featureless similarities, Journal of Data Mining and Knowledge Discovery, Vol. 15, No. 3, pp. 34–45 (2015).
[4] R. Karthiban, A QoS-Aware Web Service Selection Based on Clustering, International Journal of Scientific and Research Publications, Vol. 4, No. 2, pp. 23–34 (2014).
[5] H Abu Sharkh and B.C. Fung, Service-Oriented Architecture for Sharing Private Spatial-Temporal Data, Proceedings of International Conference on Cloud and Service Computing (CSC), pp. 40–47 (2011).
[6] C.L. Chen, F.S. Tseng and T Liang, An integration of Word Net and fuzzy association rule mining for multi-label document clustering, Journal of Data & Knowledge Engineering, Vol. 69, No. 11, pp. 1208–1226 (2010).
[7] S. Dasgupta, S. Bhat and Y Lee, Taxonomic clustering of web service for efficient discovery, Proceedings of the 19th ACM international conference on Information and knowledge management ACM, pp. 1617–1620 (2010).
[8] S. Dasgupta, S. Bhat and Y. Lee, Taxonomic clustering and query matching for efficient service discovery, Proceedings of IEEE International Conference on Web Services (ICWS), pp. 363–370 (2011).
[9] A.S. Sukumar, J. Loganathan and T. Geetha, Clustering web services based on multi-criteria service dominance relationship using Peano Space filling curve, Proceedings of IEEE International Conference on Data Science & Engineering (ICDSE), pp. 13–18 (2012).
[10] H. Dong, F.K. Hussain, and E. Chang, A survey in semantic search technologies, Proceedings of IEEE 2nd International Conference on Digital Ecosystems and Technologies, DEST, pp. 403–408 (2008).
[11] H. Gao, W. Stucky, and L. Liu, Web services classification based on intelligent clustering techniques, Proceedings of IEEE International Conference on Information Technology and Applications, IFITA'09, pp. 242–245 (2008).
[12] J. Wang, X. Yang and K. Long, Web DDoS detection schemes based on measuring user's access behavior with large deviation, Proceedings of Global Telecommunications Conference (GLOBECOM), pp. 1–5 (2011).
[13] N. Gholamzadeh and K. Taghiyareh, Ontology based fuzzy web services clustering, Proceedings of 5th International Conference on Telecommunications, pp. 721–725 (2010).
[14] Y.J. Lee and C.S. Kim, A learning ontology method for restful semantic web services, Proceedings of the International Conference on Web Services, pp. 251–258 (2011).
[15] H. Li, X. Xu, D. Hu, X. Qu, X. Tao and P. Zhang, Graph method based clustering strategy for femtocell interference management and spectrum efficiency improvement, Proceedings of 6th International IEEE Conference on Wireless Communications Networking and Mobile Computing (WiCOM), pp. 1–5 (2010).
[16] R. Mohana, and D. Dahiya, Optimized Service Discovery using QoS based Ranking: A Fuzzy Clustering and Particle Swarm Optimization Approach, Proceedings of IEEE Annual Computer Software and Applications Conference, pp. 452–457 (2011).
[17] T. Wen, G. Sheng, Y. Li, and Q. Guo, Research on Web service discovery with semantics and clustering, Proceedings of 6th International Conference on Information Technology and Artificial Intelligence Conference (ITAIC), Vol. 1, pp. 62–67 (2011).
[18] A. Nagy, C. Oprisa, I. Salomie, C.B. Pop, V.R. Chifu, and M. Dinsoreanu, Particle swarm optimization for clustering semantic web services, Proceedings of 10th International Symposium on Parallel and Distributed Computing IEEE, pp. 170–177 (2011).

[19] V. Paliwal, B. Shafiq, J. Vaidya, H. Xion and N. Adam, Semantics-Based Automated Service Discover, IEEE Transactions on Services Computing, Vol. 5, No. 2, pp. 260–275 (2012).

I. R. Praveen Joe is presently a research scholar in the Department of Computer Technology, MIT Campus, Anna University, Chennai, India. He has completed his MCA degree from Loyola College, Chennai and M.E degree from College of Engineering, Anna University, Chennai, India. He is put in around 15 years of experience in engineering teaching. He is an Associate Professor in the Department of Computer Science and Engineering, KCG College of Technology, Chennai. His technical areas of interests include web services and data mining.

P. Varalakshmi is an Associate Professor in the Department of Computer Technology, MIT Campus, Anna University, Chennai, India. She holds a B.E. degree from GCT, Coimbatore., M.Tech. degree from Pondicherry University and Ph.D. degree from Anna University, Chennai. Her research areas include Cloud and Grid computing, Network security, Mobile computing, Compilers, Theory of Computation and IoT. She is a recognized research supervisor in Anna University and has vast teaching and research experience with publications of repute.