VICTORIA UNIVERSITY OF WELLINGTON
Te Whare Wananga o te Upoko o te Ika a Maui

School of Engineering and Computer Science
Te Kura Mātai Pūkaha, Pūrorohiko

PO Box 600
Wellington
New Zealand

Tel: +64 4 463 5341
Fax: +64 4 463 5045
Internet: office@ecs.vuw.ac.nz

Optimization of Location Allocation of Web Service using non-dominated sorting algorithm(NSGA-II)

Boxiong Tan

January 6, 2015

Submitted in partial fulfilment of the requirements for
.

Abstract

1 Introduction

Web Services are considered as self-contained, self-describing, modular applications that can be published, located, and invoked across the Web [14]. In recent years, web services technology is becoming increasingly popular because the convenience, low cost and capacity to be composed into high-level business processes [1].

With the ever increasing number of functional similar web services being available on the Internet, the web service providers (WSPs) are trying to improve the quality of service (QoS) to become competitive in the market. QoS also known as non-functional requirements to web services, is the degree to which a service meets specified requirements or user needs [19], such as response time, security and availability. Among numerous QoS measurements, service response time is a critical factor for many real-time services, e.g. traffic service or finance service. Service response time has two components: transmission time (variable with message size) and network latency [12]. Study [12, 11] has shown that network latency is a significant component of service response delay. Ignoring network latency will underestimate response time by more than 80 percent. Since network latency is related to network topology as well as physical distance [10]. The network latency could also vary with the network topology changes. The only way to reduce the network latency is move the service to a location where has smaller network latency to the user center. Hence, the WSPs need to consider which physical location to deploy their services so that it could minimize the cost as well as ensure the QoS.

The Web service location-allocation problem is essentially a multiobjective optimization problem [2]. Because of the confliction between service quality and deployment cost. Ideally, WSP could deploy their services to each user center in order to provide the best quality. That is, the more services deployed, the better the quality and the higher cost. This problem is considered as an NP-hard due to the fact that the combinatorial explosion of the search space [16].

Very few researches [1, 15] study this problem. Both studies try to solve the problem by integer linear programming techniques. However, integer programming techniques do not scale well, so that no satisfactory results can be obtained for large-scale datasets.

Evolutionary algorithms (EAs) have been used in solving multi objective optimization problems in recent years. EAs are ideal for solving multi objective optimization problems [6], since EA works with a population of solutions, a simple EA can be extended to maintain a diverse set of solutions. With an emphasis for moving toward the true Pareto-optimal region, an EA can be used to find multiple Pareto-optimal solutions in one single simulation run [13].

Hai [8] proposed an enhanced genetic algorithm-based approach which make use of the integer scalarization technique to solve the multiobjective problem. The genetic algorithm (GA) is an EA that uses genetic operators to obtain optimal solutions without any assumptions about the search space. This algorithm solve the scalability problem in the dataset, however the integer scalarization technique [2] has some disadvantages:

1. The decision maker needs to choose an appropriate weights for the objectives to retrieve a satisfactorily solution.

2. The algorithm does not produce an uniform spread of points on the Pareto curve. That is, all points are grouped in certain parts of the Pareto front.

3. Non-convex parts of the Pareto set cannot be reached by minimizing convex combinations of the object functions.

Evolutionary multi objective optimization (EMO) methodologies on the other hand, successfully avoid the above mentioned problems and demonstrated their usefulness in find a well-distributed set of near Pareto optimal solutions [1]. Non-dominated sorting GA (NSGA-II) [4], Strength Pareto Evolutionary Algorithm 2 (SPEA-2) [3] have become standard approaches. Some schemes based on particle swarm optimization approaches [7, 9] are also important. Among numerous EA approaches, NSGA-II is one of the most widely used methods for generating the Pareto frontier. NSGA-II implements elitism and uses a phenotype crowd comparison operator that keeps diversity without specifying any additional parameters [5]. In our approach, we apply a modified version of NSGA-II since the web service location-allocation is a discrete problem.

In this paper we consider the problem faced by a WSP who has existing facilities but wishes to use the collected data to re-allocate their services in order to maximum their profit. The WSP must decide on facility locations from a finite set of possible locations. In order to make the decision, the WSP must first analyze the data which were collected from current services. The collected data should includes the records of invocation from each unique IP address. Therefore, based on these data, the WSP could summarize several customer demand concentrated at n discrete nodes [1], namely user centers. We assume the WSP has already done this step and list of user centers and candidate service deployment locations are given. In addition to decide which location to re-allocate the services, a dataset which contains the network latency between demand user center and candidate location are critical. The WSP could collect the data or use existed dataset [18, 17]. Then, the service provider could use the algorithm which proposed by this paper, to select an optimal plan based on their funds. The algorithm will produce a near optimal solution which indicate the services deployment locations with a minimum cost and best service quality. The main objectives are:

- To model the web service location-allocation problem so that it can be tackled with NSGA-II
- To develop a modified NSGA-II approach for the web service location-allocation problem
- To evaluate our approach by comparing it to a GA approach which use integer scalarization technique.

2 Problem Description

2.1 Model formulation

The problem is determine which facility locations that could maximus WSPs profit as well as ensure low network latency. Let $S = \{1, 2, ..., s\}$ be the set of services. We assume that the demand for service is concentrated at i demand nodes $I = \{1, 2, ..., i\}$. Let $J = \{1, 2, ..., j\}$ be the set of j candidate facility locations. To model the service location-allocation problem we use four matrices: service network latency matrix L, service location matrix A, service invocation frequency matrix F and cost matrix C.

The server network latency matrix $L = [l_{ij}]$, is used to record network latency from user centers to candidate locations, where l_{ij} is a real number denotes the network latency from user center i to candidate location j. These data could be retrieved from implementing a network latency experiment or using existed datasets [18, 17].

$$L = \begin{array}{c|cccc} & j_1 & j_2 & j_3 & j_4 \\ \hline i_1 & 5.09 & 2.37 & 4.01 & 3.9 \\ i_2 & 0.8 & 2.9 & 3.2 & 1.2 \\ i_3 & 2.74 & 1.2 & 5.3 & 0.95 \end{array}$$

These data could be retrieved from implementing a network latency experiment or using existed dataset [18, 17]. The service location matrix $A = [y_{sj}]$ represents the actual service location-allocation, where y_{sj} is a binary value (i.e., 1 or 0) shows whether a service s is deployed in candidate location j or not.

$$A = \begin{array}{c|cccc} & j_1 & j_2 & j_3 & j_4 \\ \hline s_1 & 0 & 1 & 0 & 0 \\ s_2 & 0 & 0 & 1 & 1 \\ s_3 & 1 & 1 & 0 & 0 \end{array}$$

The service invocation frequency matrix $F = [f_{is}]$, is used to record services invocation frequency from user centers, which f_{is} is an integer that indicate the number of invocation in a period of time from user center i to service s. e.g. 120 invocations per day from user center i_1 to s_1.

$$F = \begin{array}{c|ccc} & s_1 & s_2 & s_3 \\ \hline i_1 & 120 & 35 & 56 \\ i_2 & 14 & 67 & 24 \\ i_3 & 85 & 25 & 74 \end{array}$$

The cost matrix $C = [c_{sj}]$, is used to record the cost of deployment of services from candidate locations, which c_{sj} is an integer that indicate the cost of the deployment fee from a candidate location. e.g 130 \$ to deploy s_1 from j_1.

$$C = \begin{array}{c|cccc} & j_1 & j_2 & j_3 & j_4 \\ \hline s_1 & 130 & 80 & 60 & 68 \\ s_2 & 96 & 52 & 86 & 78 \\ s_3 & 37 & 25 & 54 & 46 \end{array}$$

Consider the following key modeling assumptions:

1. The new WSP decides where to locate his facilities regardless if there is existed functional similar services from other WSPs.
2. This choice is made only consider two factors: total network latency and total cost.
3. We assume a fixed customer allocation policy for WSPs. In practice, Web Services typically offer clients persistent and interactive services, which often span over multiple sessions. Therefore, a dynamic reallocation scheme is not practical as it may disrupt the continuity of the services.

2.2 Discreted NSGA-2 algorithm

NSGA-2 belong to the larger class of evolutionary algorithms (EAs), which generate approximate solutions to optimization and search problems by using techniques inspired by the principles of natural evolution: selection, crossover and mutation.

The steps involved in the solution of optimization problem using NSGA-II are summarized as follows.

- Population initialization
- Non-dominated sort
- Crowding distance
- Selection
- Genetic operators
- Recombination and selection

2.3 Chromosome Representation

In our problem, the chromosome is the service location matrix A that we mentioned in the previous section.

2.4 Chromosome Representation

The objective functions of this entire problem are following:

- Minimize the total cost of services. n is the number of service, m is the number of candidate location.
- Minimize the network total latency of the services.

3

4

References

[1] ABOOLIAN, R., SUN, Y., AND KOEHLER, G. J. A locationallocation problem for a web services provider in a competitive market. *European Journal of Operational Research 194*, 1 (2009), 64 – 77.

[2] CARAMIA, M. Multi-objective optimization. In *Multi-objective Management in Freight Logistics*. Springer London, 2008, pp. 11–36.

[3] DEB, K., MOHAN, M., AND MISHRA, S. Evaluating the &epsi;-domination based multi-objective evolutionary algorithm for a quick computation of pareto-optimal solutions. *Evol. Comput. 13*, 4 (Dec. 2005), 501–525.

[4] DEB, K., PRATAP, A., AGARWAL, S., AND MEYARIVAN, T. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *Evolutionary Computation, IEEE Transactions on 6*, 2 (Apr 2002), 182–197.

[5] DEB, K., SUNDAR, J., N, U. B. R., AND CHAUDHURI, S. Reference point based multi-objective optimization using evolutionary algorithms. In *International Journal of Computational Intelligence Research* (2006), Springer-Verlag, pp. 635–642.

[6] DESAI, S., BAHADURE, S., KAZI, F., AND SINGH, N. Article: Multi-objective constrained optimization using discrete mechanics and nsga-ii approach. *International Journal of Computer Applications 57*, 20 (November 2012), 14–20. Full text available.

[7] ELHOSSINI, A., AREIBI, S., AND DONY, R. Strength pareto particle swarm optimization and hybrid ea-pso for multi-objective optimization. *Evol. Comput. 18*, 1 (Mar. 2010), 127–156.

[8] HUANG, H., MA, H., AND ZHANG, M. An enhanced genetic algorithm for web service location-allocation. In *Database and Expert Systems Applications*, H. Decker, L. Lhotsk, S. Link, M. Spies, and R. Wagner, Eds., vol. 8645 of *Lecture Notes in Computer Science*. Springer International Publishing, 2014, pp. 223–230.

[9] HUANG, V. L., SUGANTHAN, P. N., AND LIANG, J. J. Comprehensive learning particle swarm optimizer for solving multiobjective optimization problems: Research articles. *Int. J. Intell. Syst. 21*, 2 (Feb. 2006), 209–226.

[10] HUFFAKER, B., FOMENKOV, M., PLUMMER, D., MOORE, D., AND CLAFFY, K. Distance Metrics in the Internet. In *IEEE International Telecommunications Symposium (ITS)* (Brazil, Sep 2002), IEEE, pp. 200–202.

[11] JAMIN, S., JIN, C., KURC, A., RAZ, D., AND SHAVITT, Y. Constrained mirror placement on the internet. In *INFOCOM 2001. Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE* (2001), vol. 1, pp. 31–40 vol.1.

[12] JOHANSSON, J. M. On the impact of network latency on distributed systems design. *Inf. Technol. and Management 1*, 3 (July 2000), 183–194.

[13] KANAGARAJAN, D., KARTHIKEYAN, R., PALANIKUMAR, K., AND DAVIM, J. Optimization of electrical discharge machining characteristics of wc/co composites using non-dominated sorting genetic algorithm (nsga-ii). *The International Journal of Advanced Manufacturing Technology 36*, 11-12 (2008), 1124–1132.

[14] RAN, S. A model for web services discovery with qos. *SIGecom Exch. 4*, 1 (Mar. 2003), 1–10.

[15] SUN, Y., AND KOEHLER, G. J. A location model for a web service intermediary. *Decis. Support Syst. 42*, 1 (Oct. 2006), 221–236.

[16] VANROMPAY, Y., RIGOLE, P., AND BERBERS, Y. Genetic algorithm-based optimization of service composition and deployment. In *Proceedings of the 3rd International Workshop on Services Integration in Pervasive Environments* (New York, NY, USA, 2008), SIPE '08, ACM, pp. 13–18.

[17] ZHANG, Y., ZHENG, Z., AND LYU, M. Exploring latent features for memory-based qos prediction in cloud computing. In *Reliable Distributed Systems (SRDS), 2011 30th IEEE Symposium on* (Oct 2011), pp. 1–10.

[18] ZHENG, Z., ZHANG, Y., AND LYU, M. Distributed qos evaluation for real-world web services. In *Web Services (ICWS), 2010 IEEE International Conference on* (July 2010), pp. 83–90.

[19] ZHOU, J., AND NIEMELA, E. Toward semantic qos aware web services: Issues, related studies and experience. In *Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on* (Dec 2006), pp. 553–557.