Peng LI[1], Junhao WEN[1]*, Xue LI[2]

Chongqing University (1), the University of Queensland (2)

SNTClus: A Novel Service Clustering Algorithm based on Network Analysis and Service Tags

Abstract. *The conventional service clustering methods are based on analysis of service functional and non-functional properties but ignore the hidden relations behind services contexts. In this paper, we propose a new service clustering algorithm SNTClus which based on heterogeneous service network analysis and service tags, after service ranking and clustering processes, services are clustered and ranked concurrently and K ranked service clusters are returned as output of clustering. The experiments show the feasibility and the accuracy of our method.*

Streszczenie. *W artykule zaproponowano nowy algorytm klasteryzacji usług SNTClus, oparty na analizie heterogenicznej usług sieciowych i etykiet. W pierwszej fazie dokonuje on oceny usług i procesu klasteryzacji, po czym usługi i klastry są jednocześnie kategoryzowane. Jako wynik otrzymywane są klastry kategorii K. (**SNTClus: nowy, bazujący na analizie sieci i etykietach usługowych algorytm klasteryzacji usług**).*

Keywords: Service Clustering; Network Analysis; Heterogeneous Service Network; Service Tag
Słowa kluczowe: klasteryzacja usług, analiza sieci, heterogeniczna sieć usługowa, etykiety usługowe.

1. Introduction

With the increase of Web Services, how to find the most suitable services from the huge amounts of diversiform services is one of the urgent problems of service computing. Service clustering provides a solution for service selection which can discovery the potential service categories and narrow the scope of service selection. The traditional service clustering methods are mostly based on WSDL descriptions which realize service clustering according similarities between service features in description files such as name, content, type and message [1,2,3].With the increase of complexity on service structures and contexts, WSDL based service clustering can't catch the dynamic relations between services and the related context participants such as providers and requesters, so this kind of clustering method can't meet the requirements of high accurate service clustering.

Recently, with the development of information network analysis, some new researches give us suggestions on service network and clustering. RankClus[4] is a clustering algorithm based on two-types heterogeneous information network analysis, in the clustering process clustering and ranking are enhanced with each other. NetClus[5] is another network based clustering algorithm which the network object types to three. Both RankClus and NetClus consider the network relations between different types of objects and extract the hidden categories behind information objects. As a special type of information, web services can also clustered based on service network analysis [6, 7]. Affected by the development of social network and community share technologies, some service registers or search engines have imported service tags as the additional description option beside WSDL files. For published services, different service tags mean different semantics of services, Service tags can give the keywords on functions, properties, features of services which can be used as an important computing basis in service clustering process. In [8], authors proposed a new service clustering method WTCluster in which both WSDL documents and tags are utilized for web service clustering.

In this paper, combine researches of clustering and service tags, we propose a service clustering method SNTClus which considers network analysis and service tags. The contributions of our paper are shown as:(1)We propose heterogeneous service network model which describes network structure between services, providers, requesters and tags,(2)We propose heterogeneous service network analysis based service clustering algorithm SNTClus according to service ranking functions and probability ranking model,(3)Experiments on real services show the performance of our method.

2. Heterogeneous service network model

In our research, considering service tag and network structure of services, we propose a new kind of heterogeneous network to find the potential relations between services and other related types of objects such as providers, requesters and service tags.

Definition 1: Heterogeneous service network. Define $HSN =< H, W >$ is a heterogeneous service network, where $H = S \cup P \cup R \cup T$ is the nodes set of the network HSN, S is the service nodes set, P is the service provider nodes, R is the service requester nodes set, T is the set of service tag nodes; W is the $n \times n$ relation matrix, in which for any pair of nodes $h_i, h_j \in H$, $W(i, j)$ denotes the link weight between the two nodes, the value is 0 if there is no link between h_i and h_j. For simple, we use sub-matrixes W_{SP}, W_{SR} and W_{ST} to denote the relations of network between services and any other three types of nodes. Fig.1. is a heterogeneous network example.

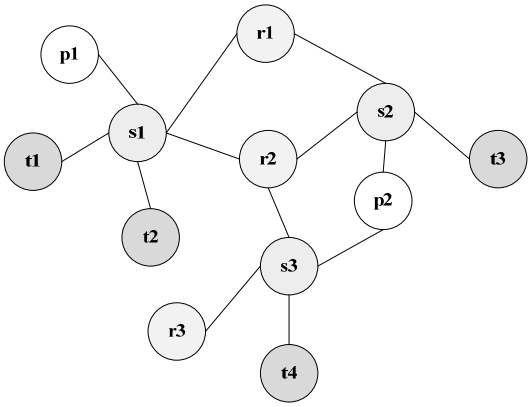

Fig.1. Heterogeneous service network example

In the example shown in Fig.1, the nodes are made up of twelve nodes: three services(s1,s2,s3),two providers (p1,p2), three requesters (r1,r2,r3) and four tags (t1,t2,t3,t4).The three relation matrixes of network are shown as:

$$W_{SP} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}, W_{SR} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}, W_{ST} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

3. SNTClus algorithm

In service computing environment, service publications are generally to solve some certain problems in some related areas; different services belong to diverse categories according to their build-in functions. Besides services, service providers and requesters are also areas related, people belonged to some special areas more possible to provide or request the area related services. With the development of tagging technology, service tags can be used to give more complete descriptions to services, service tags are also in some special areas, services in same area are more likely to be tagged by common tags. Based on the area-related feature, we propose a new kind of service clustering algorithm SNTClus. The clustering process in SNTClus is based on the service network ranking functions and ranking model, the process includes:

(1) Service Network Extraction: Extract the heterogeneous network $HSN =< S \cup P \cup R \cup T, W >$ from services request history.

(2) Network Partition: In the beginning of clustering, we random divide the extracted network to K sub-networks and each sub-network as a service cluster, the whole service network can denote as K clusters $HSN = \{C_1, C_2, ..., C_K\}$.

(3) Service Ranking: Ranking distribution in each cluster can be computed based on the network ranking model, the higher one service ranks in a cluster.

(4) Clustering Measurement and Adjusting: After service ranking, we can get the K-dimensional vector description for each service, each dimension denotes the ranking distribution in one cluster; we can also get the centroids for the K clusters by calculating the average of all the services' vector value in each cluster. We can compute the similarities between each service and every cluster, for each service, assign the cluster with the maximum similarity as the new cluster which the service belongs to.

(5) Iterative Clustering: Repeat the process from (3) to (4) until the clustering result return to a stable status.

(6) Finish and Output: Finish service clustering process and output the K service clusters.

In our research we believe that the ranking of services should high associated with the related linked nodes such as service tags, providers and requesters. In order to give a reasonable ranking model for service clustering, we define ranking functions for different nodes: service tags (Tag ranking function), service providers and service requesters (Indirect Ranking Function).

Definition 2. Tag ranking function. In a heterogeneous service network $HSN =< H, W >$, function f_T denote the ranking distribution of service tags in HSN, where for arbitrary tag node $t \in H$, the ranking of t in network HSN is defined as:

$$f_T(t \mid HSN) = \frac{\sum_{s \in LS(t)} W_{ST}(s,t)}{\sum_{t' \in T} \sum_{s' \in LS(t')} W_{ST}(s',t')} \tag{1}$$

$f_T(t \mid HSN)$ is the ranking distribution of tag node t in network HSN, $LS(t)$ denotes the linked services of tag t, T means all the tag nodes in HSN, W_{ST} is the relation matrix between services and tags, $W_{ST}(s,t)$ denotes the weight of links between service node s and tag node t.

The relations between providers and requesters are not direct links but indirect connects through the centre service nodes. In the research, we find that there is a propagation relation between providers ranking and requesters ranking which means ranking for one kind of nodes will be affected by another kind of nodes through the linked centre services. The details of ranking function are shown as follows.

Definition 3. Indirect ranking function. In $HSN =< S \cup P \cup R \cup T, W >$, the indirect ranking distributions for service providers and requesters can defined as:

$$\begin{aligned} f_P(P \mid HSN) &= W_{PS} W_{SR} f_R(R \mid HSN) \\ f_R(R \mid HSN) &= W_{RS} W_{SP} f_P(P \mid HSN) \end{aligned} \tag{2}$$

In the ranking function, $f_P(P \mid HSN)$ is the ranking distribution for service providers in HSN, $f_R(R \mid HSN)$ is the ranking distribution of service requesters in HSN. The ranking of providers and requesters have propagation between each other, in the process of indirect ranking, service nodes play a bridge role between providers and requesters. The solutions for ranking functions of providers and requesters can be calculated literately.

In our research, we user probability model to define the ranking model of services, the higher probability distribution one service ranks in a cluster, more possible the service belongs to the cluster. In our research, the ranking model includes two parts: priori probability ranking model and posterior probability ranking model.

In service context environment, according to analysis based on heterogeneous service network, ranking of services is high related to the linked providers, requesters and service tags. For one service, if all the linked nodes have higher ranking, the service should also has a higher ranking, according to this basic idea, we define the priori probability ranking model for service:

$$P_{prio}(s \mid HSN) = \prod_{h \in L(s)} f(h \mid HSN)^{W(s,h)} \tag{3}$$

$P_{prio}(s \mid HSN)$ is the priori probability ranking of service node s in network HSN, $L(s)$ is the nodes set which linked to node s, $f(h \mid HSN)$ is the ranking distribution of linked node h in network HSN which can be calculated by ranking functions we defined , (Eq. (1), Eq. (2)), $W(s,h)$ is the link weight between node s and h.

In order to improve the accuracy of service ranking, we want to get the posterior probability for each service. Among the methods of probability theory, Bayesian rule is the most popular one, in our research, we also use Bayesian rule to define the posterior probability ranking model:

$$P_{post}(C_i \mid s) = P_{prio}(s \mid C_i) \times P(C_i) \tag{4}$$

$P_{prio}(s \mid C_i)$ is the posterior probability of service s in cluster C_i, $P_{prio}(s \mid C_i)$ is the priori probability of service s, $P(C_i)$ is the possible size of cluster C_i, in our research, we use EM algorithm to get the optimum value of $P(C_i)$. After computing of posterior probability, a service is expressed as $Vec(s) = (P_{\text{post}}(s \mid C_1), P_{\text{post}}(s \mid C_2), ..., P_{\text{post}}(s \mid C_K))$ which can be used in clustering measure and adjusting processes.

Based on the service ranking model, we can get the K-dimensional vector format of every service, each dimension $P_{\text{post}}(s \mid C_i)$ denotes the ranking distribution in cluster C_i. After we get vector descriptions for all the services we can get the centroids for the K clusters by calculating the average of all the services in each cluster:

$$Center(i) = \frac{1}{|C_i|} \sum_{j=1}^{|C_i|} Vector(s_j) \tag{5}$$

After computing of each centroid for K clusters, we can calculate the similarities between each service and each cluster centroid based on Cosine Similarity method, services will be assigned to the most similar cluster. The process of service clustering and ranking is an iteration loop until service clusters don't change any more. The output of clustering based on SNTClus is K ranked service clusters.

4. Experiments

We use services from Titan [8, 9] as our data set for experiment. In our dataset, we manually select four types of services: Stock, XML, SOAP and SMS which includes 426 services with 80 service tags, 56 providers, 630 program-created requesters and 5,556 request records

In our experiment, we set clusters number K=4, after clustering of SNTClus, services are clustered into four categories as shown in Table 1.

Table 1. Service clustering result

Cluster1	Cluster2	Cluster3	Cluster4
SendSMS	StockInfoWS	XMLFeed	SoapBox
SMS_WS	StockTicker	ndfdXML	invSoap
SMSGate	StockQuote	MarketXml	SoapReceive
PremSMS	getstockbydonor	XML2PDF	SoapMonitor
SMSPro	Stock_History	XmlRouter	SoapCaller

As shown in Table 1, services are clustered into four categories, Cluster1 is SMS services, Cluster2 is Stock services, Cluster3 is XML services and the last one is SOAP services.

In order to validate the clustering accuracy of our method, we use Precise, Recall and F-measure to evaluate by experiment. The accuracy functions are shown as:

$$(6)\quad Precise = \frac{Hit(\omega_i)}{Hit(\omega_i) + Error(\omega_i)},\ Recall = \frac{Hit(\omega_i)}{Hit(\omega_i) + Miss(\omega_i)}$$

$$F - measure = \frac{2 \times Precise \times Recall}{Precise + Recall}$$

$Hit(\omega_i)$ is the number of correct ones in cluster ω_i, $Error(\omega_i)$ denotes the wrong number in ω_i, $Miss(\omega_i)$ means the number of services should be in ω_i but lost, F-measure is the trade-off of precise and recall.

In SNCTClus, we use service tags as one kind of nodes in service network, in order to find the importance of tags in service clustering, we random choose tags from dataset and create nine groups of experiments(0 to 80), final results are average accuracy of all clusters, as shown in Fig.2.

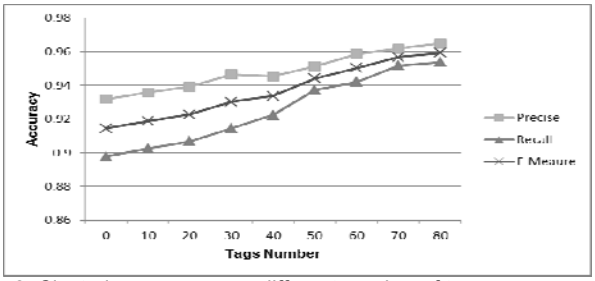

Fig.2. Clustering accuracy on different number of tags

From Fig.2 we can see, with the increase of service tags, the Precise, Recall and F-measure values will be higher, service tags can help on service clustering.

In experiments, we want to compare our method with other methods of service clustering include WTCluster [8] and SNTClus method without considering service tags. We design three groups of experiments to run three methods separately and the clustering accuracies are shown in Fig.3.

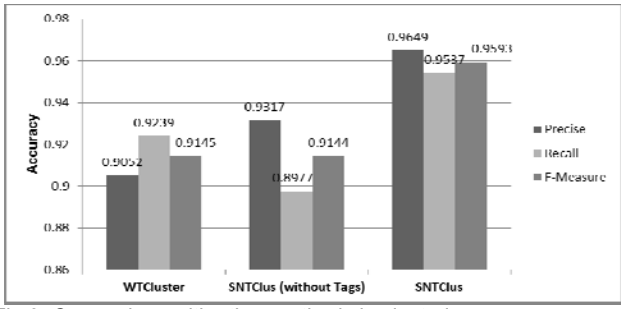

Fig.3. Comparison with other methods in clustering accuracy

From Fig.3 we can get that, compared with other two methods, SNTClus has the highest Precise (0.9649), Recall (0.9537) and F-measure (0.9593) on clustering accuracy.

5. Conclusion

In this paper, considering service tags and network analysis, we propose a new service clustering method SNTClus which extracts the potential heterogeneous service network from service usages history, concurrently makes ranking and clustering on service network and return K ranked service clusters. In our research, based on network analysis, we provide two types of ranking functions for tags, providers and requesters; the ranking model in clustering process includes prior and posterior probability ranking models. Experiments on real services show the effective and accuracy of our method. There are still some problems such as prediction of possible cluster number K, clustering performance on large number of services. In future, we will do some continuous works on service clustering and recommendation.

Acknowledgments

This work is supported by China Natural Science Foundation Project No.61075053 and Fundamental Research Funds for the Central Universities Project No.CDJZR10090001.

REFERENCES

[1] Liu W., Wong W., Web service clustering using text mining techniques, *IJAOSE*, (2009), 6-26

[2] Elgazzar K., Hassan A.E., Martin P., Clustering WSDL Documents to Bootstrap the Discovery of Web Services, *In ICWS*, (2010), 147-154

[3] Liang Q., Li P., Hung P.C.K., Wu X, Clustering Web Services for Automatic Categorization, *In IEEE SCC*, (2009), 380-387

[4] Sun Y., Han J., Zhao P., Yin Z., Cheng H., Wu T., RankClus: integrating clustering with ranking for heterogeneous information network analysis, *In EDBT*, (2009), 565-576

[5] Sun Y., Yu Y., Han J., Ranking-based clustering of heterogeneous information networks with star network schema, *In KDD*, (2009), 797-806

[6] Wang H.,Feng Z.,Sui Y.,Chen S.,Service network: An infra-structure of web services, *IEEE International Conference on Intelligent Computing and Intelligent Systems*,3(2009),303-308.

[7] E H., Song M., Song J., Li Y., Ren Z., The Research of Service Network Based on Complex Network, *2010 International Conference on Service Sciences (ICSS)*, (2010), 203-207.

[8] Chen L., Hu L., Zheng Z., Wu J., Yin J., Li Y., Deng S., WTCluster: Utilizing Tags for Web Services Clustering, *In ICSOC*, (2011), 204-218

[9] Wu J., Chen L., Xie Y., Zheng Z., Titan: a system for effective web service discovery, *In WWW*, (2012), 441-444

***Authors**: Peng Li, College of Computer Science, Chongqing University, Shapingba Road 174, 400030 Chongqing, China, E-mail:pengli@cqu.edu.cn;Junhao Wen(Corresponding author), College of Computer Science, Chongqing University,400030 Chongqing, China, E-mail: jhwen@cqu.edu.cn; Xue Li, School of Information Technology and Electrical Engineering, University of Queensland,4072 Brisbane, Australia,E-mail: xueli@itee.uq.edu.au.*