様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ３０ 年　６ 月 １５ 日現在

機関番号：２１６０２

研究種目：基盤研究(C)（一般）

研究期間：2015～2017

課題番号：１５Ｋ００４２８

研究課題名（和文）リンクドサービスネットワークのビックデータ上でのウェブサービス発見と連携

研究課題名（英文）Discovery and Composition of Web Services on Big Data of a Linked Services Network

研究代表者

Ｐａｉｋ　Ｉｎｃｈｅｏｎ(Paik, Incheon)

会津大学・コンピュータ理工学部・教授

研究者番号：７０３３６４７８

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：本研究はHadoopビックデータ下部の上、Global Social Service Network (GSSN)を生成し、その上、大量のサービスの発見をより早くすることである。
本研究の寄与として、初め、Map-Reduce GSSN(MR-GSSN)と呼ぶ、Hadoopの上動く並列計算のためのM-Rアルゴリズムを開発して18台のノードの上、30倍の早いスピードで計算した。次には、MR-GSSNの上、サービスの発見の性能を効果的に評価する方法を開発した。最後は、その他の応用として、分散コンピューティング環境の下、より効果的にタスクを割り当てをするアルゴリズムを開発した。

研究成果の概要（英文）：The objective and plan of this research is to develop distributed algorithm and system to discover services on Global Social Service Network (GSSN) on a distributed big data infrastructure and its evaluation and application. The contribution of this research is as follows. First, a novel algorithm, called Map-Reduce Global Social Service Network (MR-GSSN), to generate large service network, has been developed on Hadoop cluster with 18 nodes. We evaluated service discovery performance based on MR-GSSN, and it shows almost same result as that of GSSN with 30 times speed up. Second, in this research, we proposed a new evaluation matric for service discovery on MR-GSSN has been developed. Third, as an application of big data infrastructure, a task allocation algorithm on big data infrastructure and its evaluation has been proposed.

研究分野：サービスコンピューティング

キーワード：Service Discovery　Service Network　Big Data Infrastructure　Map-Reduce Algorithm　Task Allocation

１．研究開始当初の背景

Web services have been considered to have a tremendous impact on the web, as a potential solution for supporting a distributed service-based economy on a global scale. However, despite outstanding progress, uptake on a Web scale has been significantly less than initially anticipated. On the one hand, the number of services available on the web is far less than the expectation. The number of services in commercial field is far smaller than that of Web documents, and other academic enquiries into crawling and indexing Web services on the Web have found far smaller numbers of services. For example, the number of publicly available services contrasts significantly with the billions of Web pages available and, interestingly, is not significantly greater than the 4,000 services estimated to be deployed internally within Verizon. On the other hand, the handicap of service discovery and automatic service composition results in a lack of applications for using the services in the computer industry. Most services published on the web are never used; only about 7% of services on the web have ever been discovered, composed or invoked. From a technological perspective, the reasons can be mainly summarized into the following:

(1) All the approaches based on current service description consider only services as functional isolated islands without any links to related services, which unfortunately is hampering service discovery and service composition.

(2) Services are considered keyword or semantic matching only in terms of their own functional and nonfunctional properties through their life-cycle; and the service's social activities, defined as engaging in significant social interaction with peer services via network models, are ignored. This approach does not guarantee the quality of service discovery.

In our previous research[1] on the service discovery on linked service principle, we have constructed social linked service model based on linked data principle and first stage of global social service network concept. We could get improved performance for service discovery comparing other discovery approaches. In this research, we suggest more realistic and efficient network model for better discovery and composition of services based on complex network theory, elaborated properties for link quality, lightweight ontology principle, and inducing active participation of the global space. Finally, publication of services on global social service network (GSSN) requires a lot of computation to obtain several social link properties and to calculate traversing network to find a required services or to extract composition workflow on a graph of the network. A novel Big Data infrastructure for the GSSN will be investigated.

２．研究の目的

In this research, in order to address the aforementioned issues, a methodology to drive an innovation from the service islands isolated into global social service network by using complex network theories will be investigated on linked social service-specific principles for supporting better service discovery and composition based on Big Data infrastructure. In GSSN, services described in light-weight ontologies are published by interlinking to related services from different sources functionally across the Web and in turn be linked to from external services functionally so that services consider not only their own functional and nonfunctional detail but also service's social activities on Big Data infrastructure. Detailed research objectives are as follows:

1) Algorithm to recommend social services using the quality of social link described in "*linked social service-specific principles*" and propose a novel platform to construct global service social network to connect distributed services and analyze GSSN using complex network theories will be developed together with <K,V> scheme and Map-Reduce operation for the algorithm on Hadoop.

2) Approach to enable exploitation of GSSN, providing *linked social service* as a service will be investigated, and exploitation algorithm will be converted to Map-Reduce operation.

3) Evaluations for the global social service network and services discovery and composition will be done.

３．研究の方法

(1) SOAS Model with K-V Construction for Map-Reduce, Link Quality and Basic Parameter Formulation for GSSN to Provide Better Discovery on Big Data

There are three important sub-topics: 1) SOAS domain and properties definition using lightweight ontology model considering <K,V> scheme 2) Formulation of link property to be used for service link characteristics using Map-Reduce 3) Basic network model construction

1) SOAS domain and properties definition: Simple RDF(S) integration ontology based on the principle of minimal ontological commitment considering <K,V> scheme. Beside basic service description (IOPE), social service links and their detailed properties can be described.
2) Formulation of link quality: As a most important and basic formula to characterize the GSSN to be constructed.

Basic GSSN construction: A global space for social services registration will be constructed. The space will be based on scale-free network. Few ten thousands services from OWL-S and SAWSDL test set, Seekda.com, and ProgrammableWeb.com will be published on the space by the developed algorithm.
The services will be published on the GSSN with scale-free network using link quality obtained at the stage 2 as below:

(2) Improving Link Quality Calculation and Exploitation of GSSN for Better Composition on Big Data Infrastructure

In the second year, the basic formulation with minimal parameter set for calculating link quality and the previous GSSN will be improved. And methodologies of exploiting the GSSN to find related service will be developed.

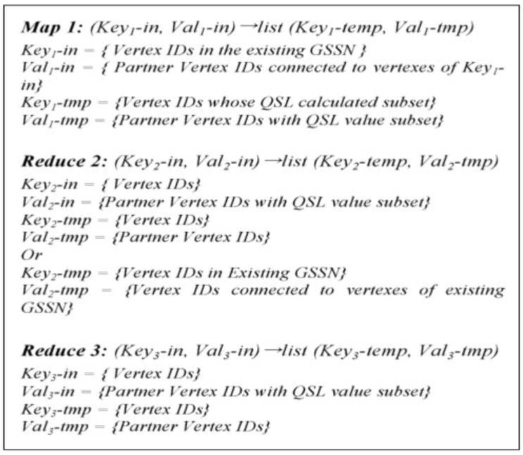
1) Improving the link quality formulation: Calculation of Q_{DSR}(R, T) is affected by service similarity by data correlation, term similarity, and ontology matching. More complete service similarity will be devised.
2) Methodology of Exploiting GSSN for Services Composition: To develop algorithms to map the GSSN into a service cluster network to reduce the search space, and a quality-driven composition approach is proposed to enable exploitation of the service cluster network, providing workflow as a service. The algorithm will be computed by Map-Reduce operation on Hadoop cluster.
3) GSSN service construction and evaluation: More complete and improved GSSN system will be constructed, and experiment of publication of services on the GSSN will be conducted on distributed Big Data Infrastructure. Also, performance comparison with Big Data approach will be conducted.

４．研究成果

There are mainly three research achievement as follows. First, an efficient Map-Reduce algorithm (MR-GSSN) to construct GSSN on Hadoop big data infrastructure has been developed.
Second, a novel service discovery performance measure on MR-GSSN has been developed.
Third, an efficient algorithm to allocate tasks on heterogeneous distributed big data centers.

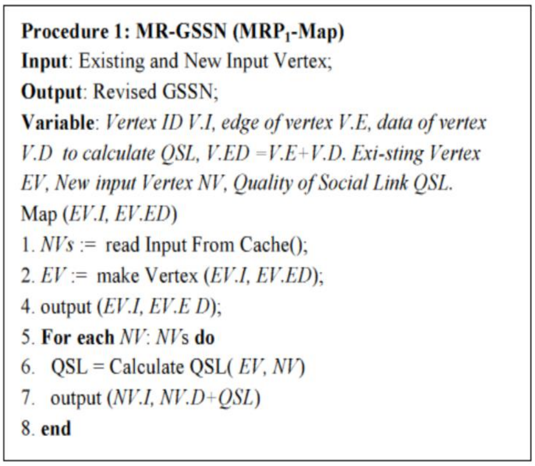
(1) Map-Reduce Algorithm for GSSN on Big Data Infrastructure

- Key-Value creation for Map-Reduce



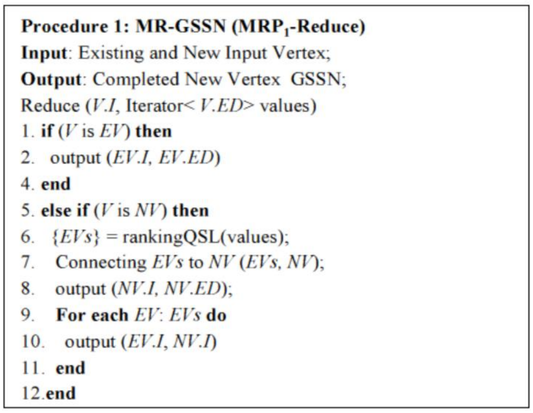
```
Map 1: ($Key_1$-in, $Val_1$-in) →list ($Key_1$-temp, $Val_1$-tmp)
$Key_1$-in = { Vertex IDs in the existing GSSN }
$Val_1$-in = { Partner Vertex IDs connected to vertexes of $Key_1$-in}
$Key_1$-tmp = {Vertex IDs whose QSL calculated subset}
$Val_1$-tmp = {Partner Vertex IDs with QSL value subset}

Reduce 2: ($Key_2$-in, $Val_2$-in) →list ($Key_2$-temp, $Val_2$-tmp)
$Key_2$-in = { Vertex IDs}
$Val_2$-in = {Partner Vertex IDs with QSL value subset}
$Key_2$-tmp = {Vertex IDs}
$Val_2$-tmp = {Partner Vertex IDs}
Or
$Key_2$-tmp = {Vertex IDs in Existing GSSN}
$Val_2$-tmp = {Vertex IDs connected to vertexes of existing GSSN}

Reduce 3: ($Key_3$-in, $Val_3$-in) →list ($Key_3$-temp, $Val_3$-tmp)
$Key_3$-in = { Vertex IDs}
$Val_3$-in = {Partner Vertex IDs with QSL value subset}
$Key_3$-tmp = {Vertex IDs}
$Val_3$-tmp = {Partner Vertex IDs}
```

- Algorithm of Map in Map-Reduce 1 Phase



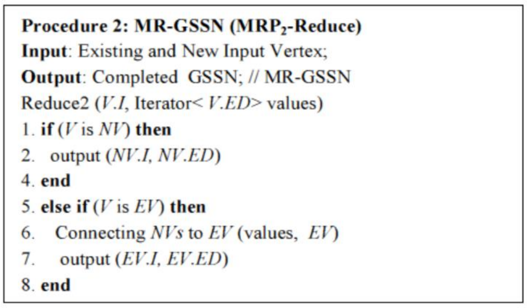
```
Procedure 1: MR-GSSN ($MRP_1$-Map)
Input: Existing and New Input Vertex;
Output: Revised GSSN;
Variable: Vertex ID V.I, edge of vertex V.E, data of vertex V.D  to calculate QSL, V.ED =V.E+V.D. Exi-sting Vertex EV, New input Vertex NV, Quality of Social Link QSL.
Map (EV.I, EV.ED)
NVs := read Input From Cache();
EV := make Vertex (EV.I, EV.ED);
output (EV.I, EV.E D);
For each NV: NVs do
    QSL = Calculate QSL( EV, NV)
    output (NV.I, NV.D+QSL)
end
```

- Algorithm of Reduce in Map-Reduce 1 Phase



```
Procedure 1: MR-GSSN (MRP$_1$-Reduce)
Input: Existing and New Input Vertex;
Output: Completed New Vertex GSSN;
Reduce (V.I, Iterator< V.ED> values)
if (V is EV) then
    output (EV.I, EV.ED)
end
else if (V is NV) then
    {EVs} = rankingQSL(values);
    Connecting EVs to NV (EVs, NV);
    output (NV.I, NV.ED);
    For each EV: EVs do
        output (EV.I, NV.I)
    end
end
```

- Algorithm of Reduce in Map-Reduce 2 Phase



```
Procedure 2: MR-GSSN (MRP$_2$-Reduce)
Input: Existing and New Input Vertex;
Output: Completed GSSN; // MR-GSSN
Reduce2 (V.I, Iterator< V.ED> values)
if (V is NV) then
    output (NV.I, NV.ED)
end
else if (V is EV) then
    Connecting NVs to EV (values, EV)
    output (EV.I, EV.ED)
end
```

Performance Evaluation

• Observation of characteristics of the MR-GSSN that is created by our distributed method. Comparing it with the GSSN that created by the existing method (by sequential addition of service node with single computer). Because our method starts form an initial node set and add service nodes by distributed calculation on multiple Hadoop nodes.

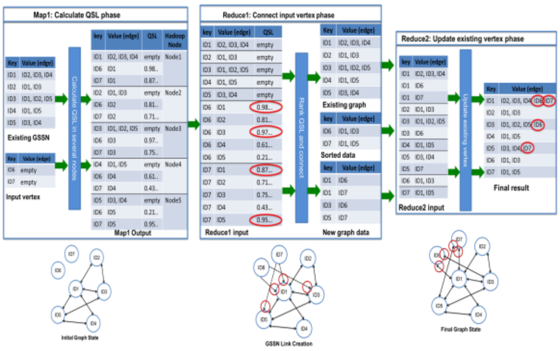


Fig. 1. Data Flow of MR-GSSN

• Observation of computation performance by our distributed method comparing with computation by a single node. Variation of computation performance on Hadoop cluster for the small size of data is large according to the computation time on each data node and other Hadoop configuration, and we made experiment considering the points.

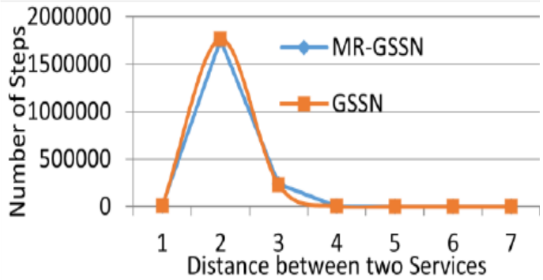


Fig. 2. Shortest Path between two services

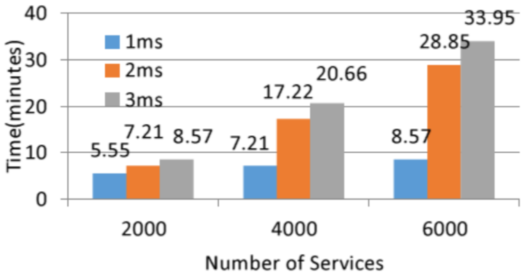


Fig. 3. Execution Time (Distributed Computing)

(2) A novel service discovery performance measure on MR-GSSN.

Service discovery and composition are challenging issue of service computing to provide value-added service. Existing approaches by keyword or ontology matching have limitations for locating realistic services discovery and composition considering non-functionality or sociality. The main reason is that approaches are based on isolated services. The isolation hinders efficient discovery and composition of services. Therefore, in the past research, they suggest social linked service network considering relationships of functional and nonfunctional properties, and social interaction based on complex network theory, where they can locate related services through sociability. However, it would be difficult to create social linked service network because services portable devices and sensors has been increasing with progress of Big Data technology. In this paper, we propose creating social linked service network to improve performance of network construction as considering distributed process on Big Data infrastructure. As for our main contributions, first, we propose an algorithm that create network graph using Map Reduce parallel programming model. Second, we evaluate performance of network graph generation and service discovery. The experimental results show that our creating network using Map Reduce approach can solve the heavy computation load for many calculations of network elements. And also, service discovery performance is very similar to that of none-distributed model.

Details for service discovery performance evaluation is as follows.
- Success Rate
Here, we evaluate success rate and existence rate of the same cluster. Success rate means that it can check whether services in GSSN connect to similar services. A service has clustering number which indicate service type. If the certain service has clustering number, we check how many services which has same clustering number connect to the certain service.
- Existence Rate of the Same Cluster
Next, we explain calculating existence rate of the same cluster. This metric is needed to check 1) degree of necessary nodes and 2) degree of recalling from service set. If service belong to a certain cluster, connected service are expected to belong to same cluster. And result of this metric can be success or fail.

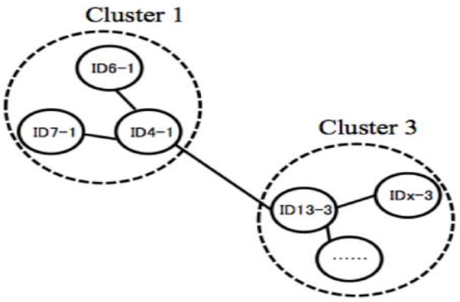


Fig. 4. Nodes and Clusters in MR-GSSN

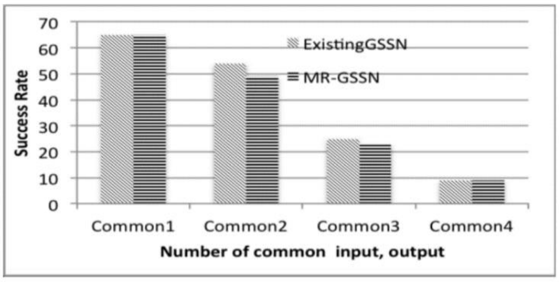


Fig. 5. Success Rate about common I/O

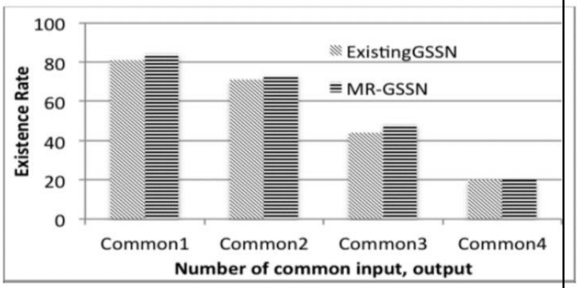


Fig. 6. Existence rate of same cluster about common I/O

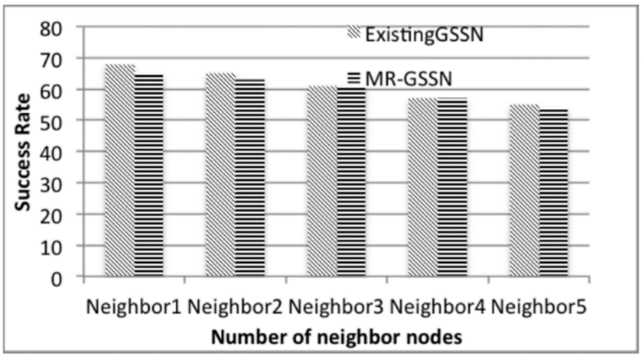


Fig. 7. Success rate about neighbor nodes

(3) An efficient algorithm to allocate tasks on heterogeneous distributed big data centers.

The virtual machine (VM) allocation problem in cloud computing has been widely studied in recent years, and many algorithms have been proposed in the literature. Most of them have been successfully applied to batch processing models such as MapReduce; however, none of them can be applied to streaming workflow well because of the following weaknesses: 1) failure to capture the characteristics of tasks in streaming workflow for the short life cycle of data streams; 2) most algorithms are based on the assumptions that the price of VMs and traffic among data centers (DCs) are static and fixed. In this paper, we propose a streaming workflow allocation algorithm that takes into consideration the characteristics of streaming work and the price diversity among geo-distributed DCs, to further achieve the goal of cost minimization for streaming big data processing. First, we construct an extended streaming workflow graph (ESWG) based on the task semantics of streaming workflow and the price diversity of geo-distributed DCs,and the streaming workflow allocation problem is formulated into mixed integer linear programming based on the ESWG.

First, an example of DC-ESWG generation for task allocation is as follows.

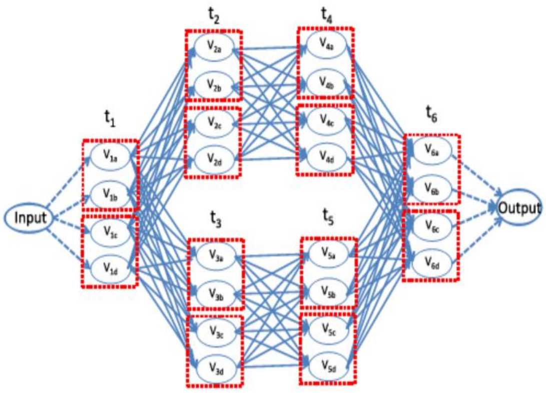


Fig. 8. Example of DC-ESWG generation

And ESWG from the previous example is as follows.

- Impact of Dc capacity and Scale of Request about cost is as follows.

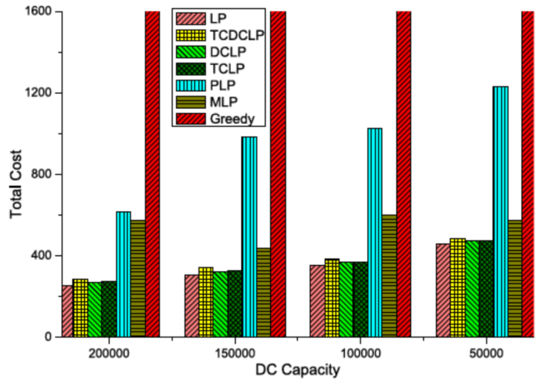


Fig. 9. Impact of DC Capacity on total cost.

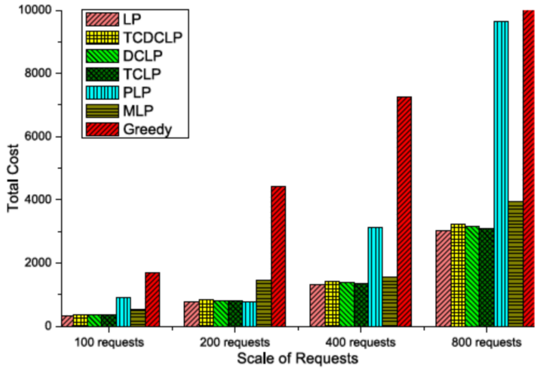


Fig. 10. Impact of scale of request on total cost.

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　2　件）

① W. Chen, I. Paik, P. C.K Hung, Constructing a Global Social Service Network for Better Quality of Web Service Discovery, IEEE Transactions on Service Computing, Volume:8 , Issue: 2, March-April, 2015, pp284 - 298.
DOI: 10.1109/TSC.2013.20

② Wuhui Chen, Incheon Paik, Zhenni Li, "Cost-Aware Streaming Workflow Allocation on Geo-Distributed Data Centers", IEEE Transactions on Computers, Vol. 66, No. 2, Feb. 2017. pp. 256-271
DOI:10.1109/TC.2016.2595579,

〔学会発表〕（計　8　件）

① Yutaka Koshiba, Incheon Paik, Wuhui Chen, Fast Social Service Network Construction using Map-Reduce for Efficient Service Discovery, Proceedings on IEEE International Conference on Service Computing 2016, San Francisco, U.S.A, June-July, 2016.

② Rupasingha A. H. M. Rupasingha, Incheon Paik, B. T. G. S. Kumara, "Calculating Web Service Similarity using Ontology Learning with Machine Learning," 2015 IEEE International Conference on Computational Intelligence and Computing Research(ICCIC), India, December 2015.

③ T. H. Akila S. Siriweera, Incheon Paik, Constraint-Driven Dynamic Workflow for Automation of Big Data Analytics based on GraphPlan,Proceedings on IEEE International Conference on Web Service 2017,Hawaii, U.S.A, June-July, 2017.

④ T. H. Akila S. Siriweera, Incheon Paik, Service Selection on BigData-Space based on Heterogeneous QoS Preferences, ICCE-Asia, Seoul, Korea, Oct. 2016.

⑤ Incheon Paik, Yutaka Koshiba, Thenuwara Hannadige Akila Sanjaya Siriweera, Efficient Service Discovery Using Social Service Network Based on Big Data Infrastructure, Proceedings of IEEE 11th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSOC 2017), Seoul, Korea, Sep. 2017.

⑥ Rupasingha A. H. M. Rupasingha, Incheon Paik, B. T. G. S. Kumara, Improving Web Service Clustering through a Novel Ontology Generation, Proceedings on IEEE International Conference on Web Service 2017, Hawaii, U.S.A, June-July, 2017.

⑦ T. H. Akila S. Siriweera, Incheon Paik, QoS and Customizable Transaction-aware Selection for Big Data Analytics, Proceedings on IEEE International Conference on Service Computing 2017,Hawaii, U.S.A, June-July, 2017.

⑧ Takeyuki Miyagi, Incheon Paik, Service Discovery on Service Network Constructed by Word Embedding, IEICE Proceedings on Technical Committee on Service Computing, June, 2017.

〔その他〕
http://ebiz.u-aizu.ac.jp

６．研究組織
(1)研究代表者
白寅天　　（PAIK, Incheon）
会津大学・コンピューター理工学部・教授
研究者番号：70336478