

Based on Automatic Correlation Keyword Grouping and Combination Based Deep Information Search Corresponding to Specific Language Big Data -- Case of Leisure Recreation

Ming-Shen Jian^{*1}, Wei-Cheng Hong¹, Sheng-Che TSAI¹, Yu-Wei CHEN¹, Chih-Ling FAN²

¹Cloud Computing and Intelligent System Lab., Dept. of CSIE, National Formosa University, Taiwan

²Dept. of Leisure and Recreation, National Formosa University, Taiwan

jianms@nfu.edu.tw

Abstract—Today, huge amount of data can be searched through the Internet. To divide the big data into several and different information groups is difficult. In this paper, the proposed Automatic Correlation Keyword Grouping and Combination Based Deep Information Search could find the further grouping information according to the specific language. By the proposed procedure, the deep searching via using simple keywords can be automatically done through the data source grouping step, Correlation Keywords Combination method, and convergence step. The implementation and verification shows that the proposed system could find the deep grouped information corresponding to different language.

Keywords—K-means, web crawler, Big Data, Hadoop, Cloud Computing

I. INTRODUCTION

Due to the community software or multimedia, the information can be published, searched, or exchanged through the Internet. Via using the search engine such as Google [1], Yahoo, etc., the similar and published information could be found corresponding to the given search keywords. Generally speaking, the data size of the search report from the search engine depends on the on demand given search keywords.

However, mostly, after searching based on the search keywords, users still need to review the content of the searched website or web pages for further information. Currently, most search engines will count the related search keywords based on the statistics. Therefore, according to the probability, the search keywords given by the users can be connected to other potential vocabularies [2, 3]. In addition, the search engine could find the required information more precisely according to more connected keywords.

Although search engine could help a lot in information searching, due to the different language structures, some less correlation of information would be provided to users. For example, the vocabularies and sentence structure of English and Chinese are various and different. The language head parameter [2-4] of Chinese and English may be different in

the location of the whole sentence. Therefore, not only the keywords comparison but also the semantic meaning should be considered when to search and to analyse the information of the web pages.

Today, there were applications or software proposed for breaking Chinese sentence and vocabularies [2-4]. However, most applications were proposed for comparing the existed vocabularies database or dictionary. If the analysed vocabulary is consisted of two independent vocabularies in Chinese, the applications will break this analysed vocabulary into two words. In other words, after wrong breaking, the analysed sentence may not be the same meaning as the original. Therefore, considering the combination of various vocabularies in Chinese is needed.

Since the content or information from the web pages or web sites may be huge, to distribute the information to different computers for parallel processing is the possible solution. Hadoop [7, 12] or Spark [17] were proposed to deal with the big data which comes from the Internet. Based on the concept of distributed system and parallel processing, all the vocabularies will be evaluated and formatted as the set: {key, value}. Each vocabulary will be evaluated as the 'key'. Then, the 'value' will be the total number of key vocabulary occurrences. By dividing the original job into tasks, distributed system can parallel process the tasks for better performance than only processing the job on single computer node.

Considering the possible gathered information from the Internet, how to find the content or data with the highest correlation about the on demand given keywords is an important issue. In addition, the discussion trend or the most discussed topics of the given keywords may be also the main theme which the users be interested in. Therefore, to find the feature of the gathered information or data is an important issue.

In this paper, automated information gathering based on web crawler is proposed. According to the proposed correlation keywords combination method, different possible

longer keywords can be found and traced back to the meaningful data or content. Finally, considering the classification results based on K-mean algorithm, the feature of the information can be obtained.

This paper is organized as follows. Section 2 introduce the problem and difficulty of vocabularies breaking in different languages. Section 3 presents the proposed correlation keywords combination method in Chinese. Section 4 gives the implementation. The conclusion is given in Section 5.

II. PAGE LAYOUT

Traditionally, the English vocabularies can be separated or identified according to the 'space' between any two words. Therefore, all the vocabularies can be counted individually. However, the structure of the 'complete vocabulary' may be different in different languages, such as Chinese. For example, in Chinese vocabulary with single word, '毛', indicates hair or feather. Another Chinese vocabulary with two words, '毛巾', indicates the towel. In other words, in some languages, a vocabulary may consists of different amount of words. In addition, the combinations of these vocabularies may starts with the same character or word. Therefore, using the 'space' to separate individual vocabulary is unavailable. More vocabulary or sentence structure should be considered.

Some online proposed methods tried to separate and identify the vocabularies by using language corpus [2-4]. By comparing the vocabularies recorded in the corpus, the matching vocabularies can be identified as the text-pattern. However, the language corpus cannot record all the vocabularies especially the newest vocabularies. To identify the unknown vocabularies is important issue which may affect the performance of the system.

A. Web Crawler

In addition, to collect the information through the website or web pages, the web crawler [5] is the possible web application service solution. The web crawler is the a program or application for data searching and collecting from the Internet. Furthermore, the web crawler application can get corresponding data to the local server for subsequent applications [6]. In other words, the web crawler can combines and collects data or information from different search engines and then presents a single report.

B. Hadoop/MapReduce

Hadoop/MapReduce which are proposed by Apache/Google are the popular applications for big data management [7,12] on the cloud computing environment. To deal with the big data, all the data will be stored according to the Hadoop Distributed File System, called HDFS [8], or the Google File System. In addition, considering the concept of parallel processing and distributed system, the data processing structure called MapReduce was proposed [9]. Furthermore, since the data is stored and distributed, the corresponding information of the data is recorded in the cloud database, called HBase [10].

Based on the HDFS [8], all the data or file will be divided into multiple partitions. In other words, the data or file will not be stored as a single file or data. In oppositions, the divided partitions, also called blocks, can be stored with less size of file or data. Since the location of individual block may be different, to record the corresponding information of the block with the same data length is difficult. Hence, the HBase [10], a noSQL type database was used to record the information or data with various data length instead of the traditional database with fixed data length. Therefore, one original file or data can be stored in different spaces with huge capacity. By using HDFS, the scattered data with huge amount of data size can be managed with higher performance.

After storing the big data, to analyse and deal with the big data with higher performance is an important issue. Based on the concept of distributed system and parallel processing, Hadoop/MapReduce [7,12] was proposed with two phases: Map and Reduce. The original big data, also called job in hadoop, will be divided into multiple tasks. All these tasks are Mapped to multiple computers for further execution or processing. After computing by each computer, a report of the task will be created and sent back to the computer which gives the task. In other words, the computer of the job owner will receive multiple reports from all computers of individual task. All the information of the reports will be presented as (key, value). Each vocabulary will be evaluated as the 'key'. Then, the 'value' will be the total number of key vocabulary occurrences. Without the doubt, the size of all the reports will be also huge. Hence, based on the Reduce phase, similar to the procedure of mapping tasks, all the report will be parallel processed in distributed system. Then, all the values with the same key will be summed or added up. Finally, only one summed report will be provided to the system user.

C. Cloud Computing

To manage and analyse the big data with the huge amount of storage space, cloud computing is considered to provide the computing resource. Through the network, the virtualized computing resource and be merged into the single computing resource pool for rapidly using [7]. According to the given requirements of cloud users, the corresponding computing resource can be integrated as the virtual machine for users. In addition, by using Docker engine, the required service which is included in the container can be implemented immediately for users without the assistance of physical hardware. When the services or virtual machines are no longer used or required, all the occupied computing resource can be returned to the resource pool. Therefore, all the virtual machines or services can be rapidly established according to the requirements of system users. In this paper, considering the dynamic computing resource requirements of big data analysis, to establish the proposed system based on the cloud environment is the possible solution.

D. K-means Clustering

To find the relationship between different data, K-means clustering algorithm was proposed. Based on the vector quantization and group analysis, different groups with

individual characteristics can be found and defined. The equation can be defined as follows:

$$J = \sum_{i=1}^k \sum_{x_j \in S_i} (x_j - \mu_i)^2 \quad (1)$$

Initially, there are k group centres chosen randomly. The chosen group centres will be named as $\mu_1 \sim \mu_k$. Then, the distance between all the data in the data space and the group centres will be evaluated. All the data x which is defined as the member of the group S_i will have the shortest path from μ_j . Finally, by finding the minimum value J , all the data x will have its own nearest group centre μ .

III. PROPOSED SYSTEM

To collect and analyse the information automatically, web crawler should be implemented for data gathering from different web search engines. Then, all the collected information should be analysed based on Hadoop/MapReduce algorithm. According to the report from the Hadoop/MapReduce algorithm, all the collected information or big data will be divided into several groups based on the K-mean. In other words, the collected data can be grouped according to the corresponding keyword list. Then, the proposed Correlation Keywords Combination Method can be used to further deep search based on the grouped keywords through the search engine. The flowchart can be shown as follows.

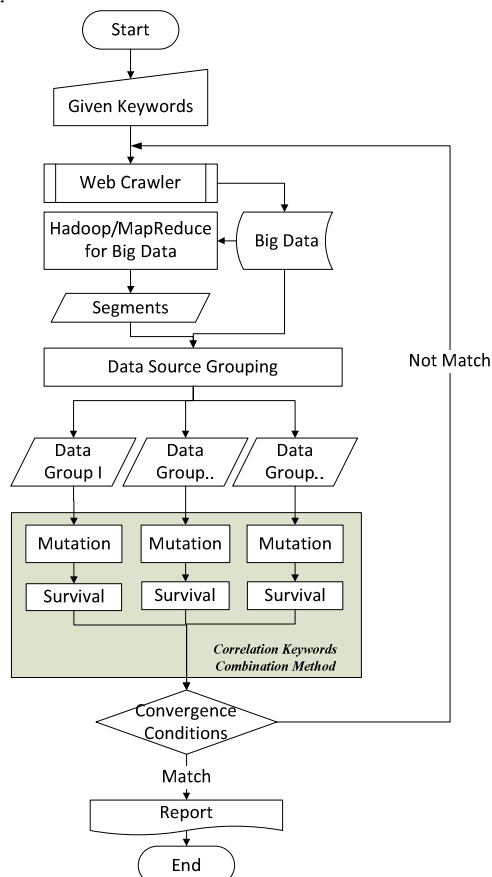


Figure 1. Flowchart of the proposed system.

A. Web Crawler

In this paper, the web crawler is established based on the search command and regular expressions of HTTP protocol [12, 13]. Through the selenium [14] and Google searching engine [1], the on demand given keywords including English and Chinese language can be searched. Based on the concept of browser automation [14], the website which contains the given keywords will be found. In other words, the web pages including the source code and content will be obtained.

After gathering the information from the web pages, to find the segment the content into suitable vocabularies is important. In this paper, considering the structure of the Chinese language, the Maximum Matching [15] program which including the ambiguity resolution rules is used, also called mmsegj written in Java language. The neighbouring vocabularies, called chunk, should follow the ambiguity resolution rules: maximum matching, largest average word length, smallest variance of word lengths, and largest sum of degree of morphemic freedom of one-character words [16]. Finally, the original content of the web pages can be divided into vocabularies.

B. Hadoop/MapReduce

After web crawler, all the web pages found according to the given keywords will be divided into vocabularies based on MMSeg. These vocabularies become the big data analysed in this paper. Based on the Hadoop/MapReduce, all the segmented and found vocabularies are represented as the keys. Then, the values which indicates the number of individual vocabulary found in whole collected web pages can be added up. In other words, after Hadoop/MapReduce procedure, the found vocabularies corresponding to the given keywords can be sorted and listed.

However, the given keywords may correspond to various themes of web pages. In addition, considering the themes of these found web pages, the pages with similar theme may use the similar even the same segments or vocabularies. Therefore, according to the popularity of the theme online, the segments or vocabularies of popular theme would be with higher visibility. The total number of segment or vocabulary occurrence in the popular theme should be much more than that in other them. Therefore, based on the sorting report from the Hadoop/MapReduce, the segments or vocabularies can be clustered according to the proposed function.

C. Data Source Grouping

According to the sorting report from the Hadoop/MapReduce, the clustered segments and vocabularies can be separated into several clusters. Based on these clusters, the original big data such as web pages or websites will be clustered.

First, each web page, called big data source, will be evaluated by vocabularies or segments matching corresponding to each clusters. Then, based on the total number of vocabularies or segments occurrence, different websites or web pages can be grouped as the big data sources with the same theme.

In other words, after searching the online information according to the given keywords, all the found data can be divided into corresponding groups related to the various themes. Since the given keywords would be corresponding to various themes, to classify the websites and web pages is important. The popularity of different themes would be also different. Hence, the correspondence between the given keywords and the themes can be found.

D. Correlation Keywords Combination Method

Although the segments and vocabularies can be found according to the MMSEG algorithm, some specific segments or vocabularies 1. be longer than three words; or 2. may be combined with more than two vocabularies. For example, the name of the store may include a name, an address or location, even a word from foreign language. Fortunately, most of specific segments or vocabularies are written or shown together in one website and web page. In addition, these specific segments or vocabularies are always the characteristic of these websites and web pages. Therefore, how to find these vocabularies becomes an important issue for the features of the data source groups classification. In this paper, the correlation keywords combination method is proposed to find the feature segment or vocabulary. The proposed method includes Mutation phase and Survival phase.

Since the data source groups can be classified according to the Hadoop/MapReduce report, the corresponding vocabularies are also found and related to the groups. Suppose that there are total g groups after Hadoop/MapReduce. Then, corresponding to the group, there are total i found segments or vocabularies, shown as Vg . In other words, the i th vocabulary or segment in group g can be represented as $Vg(i)$. Hence, the total numbers of occurrence corresponding to each vocabulary or segment can be shown as $O(Vg(i))$. However, how to select and combine the segments or vocabularies for the permutations of a longer or new segment and vocabulary creation is important.

Considering the feature segment or vocabulary would be combined from two or more composited segments or vocabularies, the number of occurrence about these composited would be quite close. Therefore, in this paper, the composition rate to select the segments or vocabularies is defined as α . Then, the segments or vocabularies which satisfy the following equation will be chosen for the possible feature segment combination, also called Mutation phase in this paper:

$$\begin{aligned} O(Vg(i)) \times \alpha &> O(Vg(j)) && \text{without combination} \\ O(Vg(i)) \times \alpha &\leq O(Vg(j)) && \text{combine as } Vg(i)Vg(j) \end{aligned} \quad (2)$$

where the $i \neq j$.

Therefore, two segments or vocabularies could new combined as two different new segments and vocabularies such as $Vg(i)Vg(j)$ or $Vg(j)Vg(i)$. Then, all these new segments will be evaluated according to the Survival phase. In the beginning, all these combined new segments and vocabularies will be used as the keys of Hadoop/MapReduce. In other words, these segments or vocabularies will be used to analyse the data of the corresponding group again.

Considering the actual content of the data source, only the feature segment or vocabulary will be reported with larger number of occurrence. Finally, comparing new segments from the same selected vocabularies, only the segment with larger number of occurrence could be survival and reserved. In other words, after Survival phase, total number of vocabularies or segments can be reduced. The survived segments or vocabularies called correlation keywords can be found and listed corresponding to each theme or group.

E. Convergence

Considering the found correlation keywords via the proposed correlation keywords combination method, the length of these survival keywords or segments will be increased. These survival correlation keywords will be used as the new search keywords for web crawler procedure again. In this paper, every new web crawler procedure is called new generation.

After new round of the generation, the new big data from web crawler can be obtained. In other words, the new reports of Hadoop/MapReduce can be provided. Since every generation will increase the length of correlation keywords, to completely match the correlation keywords will be difficult and the total number of individual correlation keyword occurrence will be decreased. Therefore, in this paper, the convergence value β is used to decide when to stop the new generation. Based on the Hadoop/MapReduce and Data Source Grouping, all the found segments or vocabularies are sorted according to the number of occurrence. Then, the segment or vocabulary with most number of occurrence in individual group can be shown as $Vg(1)$. In addition, all the websites or web pages from the web crawler can be recorded. Suppose that the total number of websites or web pages in each data source group is defined as Wg . Then, the convergence condition can be defined as follows:

$$\begin{aligned} \frac{O(Vg(1))}{Wg} &< \beta && \text{convergence} \\ \frac{O(Vg(1))}{Wg} &\geq \beta && \text{new generation} \end{aligned} \quad (3)$$

In other words, when the convergence condition is satisfied, it means that the length of the new correlation keywords are long enough and the probability of completely keywords or segments matching is low. No further correlation keyword searching is needed. In opposition, if the convergence condition is not satisfied, it means that the correlation keywords are still too popular and the length of these keywords can be increased. Finally, when the convergence condition is satisfied, the correlated key segments of the corresponding groups or themes can be found. In other words, from the initial given keywords, the proposed system could find the separated groups of key segments (more vocabularies) which are related to the corresponding themes. Furthermore, the groups of segments found can be listed according to the popularity of theme. The relationships between initial given keywords and the themes can be defined.

IV. IMPLEMENTATION

In this paper, the whole proposed system is implemented for verification. To search the websites and the web pages through the Internet, the selenium web agent is used. Through the selenium, the search results from the Google search engine can be obtained. After the proposed system, each grouped data of the leisure recreation can be given the characteristics or descriptions corresponding to the keywords or key-sentences found based on the proposed system. In other words, different characteristics corresponding to the given leisure recreation keywords can be found and divided into various data groups. Figure 2 present the cycle of the proposed system.

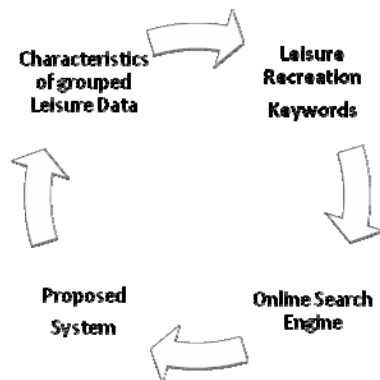


Figure 2. The cycle of the proposed system.

Then, all the online page content will be parsed to the text content as follows.

民	1068645	
工	1067530	
園	1043591	
吃	1028888	
到	1001298	
上	1001545	
本	972535	
鄉	955500	
日	954758	
鄉	949178	
苗	945861	
湖	940448	
林	925940	
宿	925899	
美	925849	
坑	925808	
西	925208	
來	909715	
高	903264	
超	894847	
黑	873929	
家	873447	
地	873141	
和	862126	
站	836054	
繪	832473	
及	820443	
花	820443	
茶	811714	
地	810660	
址	804123	
必	800438	
元	798608	
點	794047	
上	774976	
遊		773887
景	772057	
厚	771221	
店	764015	
區	764015	
行	743620	
旅	740883	
游	732469	
林	729719	
水	724352	
生	703213	
毛	699120	
手	697400	
巾	673466	

Figure 3. Some of the gathered and analyzed vocabularies of online page content

After obtaining the text content from the web pages, the proposed system integrates the mmsegj of MMSeg algorithm to find the segments or vocabularies in the text content. All

the found segments and vocabularies will be recorded as the key for the Hadoop/MapReduce procedure. The searched data from the Internet could be huge. Therefore, Hadoop/MapReduce based on Java programming language is implemented for counting the number of individual segment or vocabulary occurrence.

Since there are various websites and web pages searched according to the given keywords, all the identified segments and vocabularies should be added up. According to the number of occurrence, the segments and vocabularies are sorted from huge to few. Based on the proposed method, different vocabularies or segments will be merged or combined as the new possible correlation segments or vocabularies.

[illegible]

Figure 4. Example of proposed Correlation Keywords Combination Method

V. CONCLUSIONS

In this research, different analyzed vocabularies from Hadoop/MapReduce could be automatically merged according to the proposed Data Source Grouping method, Correlation Keywords Combination Method, and convergence method. Then, the deep grouped information can be found and defined. The system users can only give the simple keywords for managed huge data. The implementation shows the feasibility of the proposed procedure.

ACKNOWLEDGMENT

The research is supported by the project, "Ministry of Education: E-commerce and financial technology alliance centre sub-project". Thanks for the support of Cloud Computing and Intelligent System Lad. (CCIS Lab.) of National Formosa University.

REFERENCES

- [1] <http://www.google.com>
- [2] K.J. Chen & Ming-Hong Bai, "Unknown Word Detection for Chinese by a Corpus-based Learning Method," International Journal of Computational linguistics and Chinese Language Processing, 1998, Vol.3, #1, pages 27-44
- [3] Wei-Yun Ma & K.J. Chen, "A bottom-up Merging Algorithm for Chinese Unknown Word Extraction," Proc. of ACL workshop on Chinese Language Processing 2003, pages 31-38
- [4] <http://120.108.221.55/profchwu/ml/%E6%95%99%E6%9D%90%E6%96%87%E4%BB%B6%E5%88%86%E9%A1%9E%E4%B8%AD%E7%A0%94%E9%99%A2%E8%A9%9E%E6%80%A7%E5%AE%9A%E7%BE%A9.pdf>
- [5] Chandni Saini, Vinay Arora, "Information retrieval in web crawling: A survey," Advances in Computing, Communications and Informatics (ICACCI), 2016 International Conference on.
- [6] M. S. Jian, et al, "Cloud Based Agriculture Safety Inspection with Multiple Standard Sources," Proc. of IEEE ICACI, Feb. 2018.
- [7] Yu-Chen Wang, "The Structure of the Big Data Based on the Hadoop Platform", National Chi Nan University, Nantou, Taiwan, 2015.
- [8] Zhi-Kai Liao, "Fault-Tolerant Management Framework for Hadoop Distributed File System", Tamkang University, Xinbei, Taiwan, 2015.
- [9] Jia-Chun Lin, "Study of Job Execution Performance, Reliability, Energy Consumption, and Fault Tolerance in the MapReduce Framework", National Chiao Tung University, Hsinchu, Taiwan, 2015.
- [10] Man-Ning Wu, "Data Analysis Results Storage Design in HBase, Chung Yuan Christian University", Taoyuan, Taiwan, 2015.
- [11] Hadoop – Introduction, Available online : https://www.tutorialspoint.com/hadoop/hadoop_introduction.htm
- [12] Ming-Shen Jian, Yi-Chi Fang, Yu-Kai Wang, Chih Cheng, " Big Data Analysis in Hotel Customer Response and Evaluation based on Cloud," Proc. of IEEE ICACI, Feb. 2017.
- [13] Fredrik Erlandsson, Roozbeh Nia, Martin Boldt, Henric Johnson, S. Felix Wu, " Crawling Online Social Networks," Proc. of 2015 2nd ENIC, pp.9-16, Sept. 2015
- [14] <https://www.seleniumhq.org/>
- [15] Chih-Hao Tsai, "MMSEG: A Word Identification System for Mandarin Chinese Text Based on Two Variants of the Maximum Matching Algorithm," <http://technology.chtsai.org/mmseg/>
- [16] <https://www.itread01.com/articles/1476609057.html>
- [17] <https://spark.apache.org/>



Ming-Shen Jian was born in Kaohsiung City, Taiwan in 1978. He received the B.S. from the National Chiao Tung University, HsinChu, and Ph.D degrees in Computer Science and Engineering from the National Sun Yat-sen University, Kaohsiung, Taiwan in 2007.

From 2009, he was a Assistant Professor and director with the National Formosa University Cloud Computing and Intelligent System Laboratory. Since 2009, he has been an Assistant Professor with the Computer Science and Information Engineering Department, National Formosa University. He is the author of four books, more than 50

articles, and at least 15 invention patents. His research interests include IOT development and application, Big Data, Optimal Solution, Intelligent System, and Cloud Computing. He was a Secretary of the Taiwan Association of Cloud Computing.

Dr. Jian was a recipient of the IEEE sponsored international conference Paper Award in 2016, 2017, and 2018.

Chih-Ling FAN is a master degree student of Dept. of Leisure and Recreation, National Formosa University, Taiwan. Her current research interests are in the area related to Theater-style Tour Guide. She received B.S. degree at Dept. of Leisure and Recreation, National Formosa University, 2017. She joins the "Ministry of Education: E-commerce and financial technology alliance center sub-project" and Cloud Computing and Intelligent System Lad. (CCIS Lab.) from 2019.



Wei-Cheng Hong is an undergraduate student of Dept. Computer Science and Information Engineering at National Formosa University. His current research interests are in the area related to Intelligent Agriculture, Big Data, and Cloud Computing. He joins the "Ministry of Education: E-commerce and financial technology alliance center sub-project" from 2019, and Cloud Computing and Intelligent System Lad. (CCIS Lab.) from 2017.



Sheng-Che TSAI is an undergraduate student of Dept. Computer Science and Information Engineering at National Formosa University. His current research interests are in the area related to Algorithm, Big Data, and Cloud Computing. He joins the "Ministry of Education: E-commerce and financial technology alliance center sub-project" from 2019, and Cloud Computing and Intelligent System Lad. (CCIS Lab.) from 2017.



Yu-Wei CHEN is an undergraduate student of Dept. Computer Science and Information Engineering at National Formosa University. His current research interests are in the area related to Algorithm, Big Data, and Cloud Computing. He joins the "Ministry of Education: E-commerce and financial technology alliance center sub-project" from 2019, and Cloud Computing and Intelligent System Lad. (CCIS Lab.) from 2017.

