Online Courses Recommendation System based on Industry Occupation Skills Requirements

Hai-Hui Wang[1], Chalothon Chootong[2], Ankhtuya Ochirbat[3], Worapot Sommool[4], W K T M Gunarathn[5], Timothy K. Shih[6]
Department of Computer Science and Information Engineering
National Central University
Taoyuan, Taiwan
104522071@cc.ncu.edu.tw[1]
chalothon.cs@gmail.com[2]
ankhaa8@gmail.com[3]
title.wo@gmail.com[4]
tharanga.gunarathne@gmail.com[5]
timothykshih@gmail.com[6]

***Abstract*—MOOCs had bring us to a higher education with the concept of flipped classrooms, where students make use of the online studying materials such as online textbooks, video tutorials, and all sorts of documents which may take in forms of a web page, online learning platform, educational learning management systems. We see the stupendous potential of MOOCs in education. However, there has always been a problem that existed in Taiwan that is also often discussed. It is known as the gap between industry and education, which means that the students who has graduated from universities, do not always have the skills that the industries needed. We find that in most cases, students will only have some skills or knowledge about some tools that is listed from the requirements of the industries. The students have plentiful self-studying resources from the internet, we hope to encourage the students to learn and empower themselves by correctly recommend what are the most required skills of their desired occupation. Therefore, this paper proposed a clustering method that shows the results of groups of skills that are commonly needed for a particular type of job.**

Keywords—MOOCs; Job Hunting; Course Recommender; Clustering; Skills.

I. INTRODUCTION

With the rapid development of technology, human's daily life becomes more and more convenient with the growth of technology. This has changed our way of living, the changes that affected traditional learning. In traditional learning, students sit in classrooms listen to the educator on stage to learn. Nowdays, we have a better way of learning, the concept of the flipped classrooms, where students learn from online resources such as online textbooks, video, tutorials. These learning materials are often found from online learning platforms namely edX, Coursera, Udacity etc., or educational learning management systems from different university or colleges, where educators uploads their course materials to the online learning platform. Students can therefore start their learning progress with the online learning materials provided. This concept of modern learning style is also known as MOOCs, massive open online course, defined as an online course aimed at unlimited participation and open access via the web. While MOOCs becoming popular, more and more online courses will be found on web.

As the increasing amount and different categories of the online courses, it becomes crucial for the students to decide what to study from, which course to take with their limited spare time. Apart from that, another issue that is related students' learning and it is also a big problem we face in Taiwan is that students whom graduated from universities do not always have the necessary skills that the company employer required. For example, an employee requirement usually consists of a list of required skills or tools, but in the real situation, the majority of the students, whom just graduated from university, will only be able to meet a small portion of the required tools. This is also known as the gap between industry and education. However, it will take a lot of effort and time for our education system to advance to a stage for the gap to be closed. It is then impossible to change the way of education immediately.

978-1-5386-2761-7

Thus, how to solve this problem would be the main contribution of this research. It is mentioned earlier that since the raise of MOOCs, we have the necessary resources for studying, it will be very helpful to advice the students what to study in order for them to get prepared when getting a job. Therefore, in this research, the job employment data of a popular job hunting site, www.104.com.tw, is collected for the required skills or tools references, since it is the actual requirement from the industry. Then, these data is clustered in different clusters to show that some set of skills or tools are often needed for a specific type of job. In our case, this research only focuses on the Computer Science field related occupations. The result clusters will be showed in the form of a directed graph, which can represent a course map, this is to help the students to get an idea of the order to study, they can then find the related MOOC courses of the skills or tools, and self-study in their free time while completing university. This research is to help the students to enhance their own knowledge based on their occupation target in the industry, therefore increases the chances of getting the offer.

II. Related Work

There are only quite a few of the researches that are related to this concept. Some like [1], which is published in 2011, where the goal is similar as to recommend software skills for jobs in the field of Internet Technology. In this research, they proposed a solution to assist employers when preparing advertisement via identification of suitable soft skills together with its relevancy to that particular job title. Bayesian network is employed to solve this problem because it is suitable for reasoning and decision making under uncertainty. The proposed Bayesian Network is trained using a dataset collected via extracting information from advertisements and also through interview sessions with a few identified experts.

Another part that is important in this research is the clustering algorithm for the skills and tools. K-means would be the basic clustering algorithm to use. However, it would be difficult to determine the vector with this method, and the results might also be affected by the way vectors defined. [2] had a K-means clustering in the situation that the vectors can be defined in the right way. Its objective is to propose an effective clustering technique to group users' sessions by modifying K-means algorithm and suggest a method to compute the distance between sessions based on similarity of their web access path, which takes care of the issue of the user sessions that are of variable length. The basic K-means algorithm initially selects the cluster centroids randomly and finds the new cluster centroid based on the average value obtained within each cluster in each iteration. In the modified K-means algorithm, the old cluster centroid is updated by the delta amount, where, delta is nothing but the average distance value of each cluster.

Therefore, the first step is to define the similarity between the skills and tools data collected. [3] has included various ways of deciding similarity, and even combines them together for improvement. It focuses on keywords search, and designed method for measuring keywords similarity with Jaccard's, N-Gram, Vector space, Average (JNVA) and Jaccard's, N-Gram, Length, Average (JNLA) by using hybrid method; a combination of Jaccard's , N-Gram and Vector Space to make Keywords search practical. These methods are evaluated by three criterions which are precision, recall, and F-measure. The result reveals that the method for measuring keywords similarity with the application of JNVA and JNLA can successfully predict the similarity between keywords query with index words. These methods can be applied in order to develop searching engines performance especially semantic search. [4] presented one such class of item-based recommendation algorithms that first determine the similarities between the various items and then used them to identify the set of items to be recommended. Cosine-Based Similarity and Conditional Probability-Based Similarity are mainly used for calculating similarities in this paper.

There are also some other recommendation system such as [5], where they proposed a practical and effective approach, in which they first analyze the correlation between students' achievement with their employment situation, whether they have obtained an employment or not, according history data. For the data of students obtained employment, they further discover association rules from students' achievement and concrete occupations by data mining. Moreover, for a new or not obtained employment student, they recommend appropriate occupation for him or her based on the above association rules.

Some kind of filtering method was also tried during the research, the most popular would be [6], from Amazon.com. They use recommendation algorithms to personalize the online store for each customer. Rather than matching the user to similar customers, item-to-item collaborative filtering matches each of the user's purchased and rated items to similar items, then combines those similar items into a recommendation list.

As well as some other algorithms used in similar cases, [7] applied data mining techniques to social networks to help users of the social digital media to distinguish these important friends from a large number of friends in their social networks.

Other researches that also help students to get their desired occupations like [8] used developer profiles form GitHub to match the job advertisement. Employers and HR personnel who may use GitHub to learn more about a developer's skills and interests. They propose a pipeline that automatizes this process and automatically suggests matching job advertisements to developers, based on signals extracting from their activities on GitHub.

III. Proposed Method

The System Structure is shown on Figure 1. Students are free to choose a desired occupation from the job list in the system, and a course map will be formed from the complete course map that is stored in the database by using the list of required skills of this specific job. Each course on the course map will link to the search results in top online MOOCs site. When there is a new skill that appears from the analyses of sorting the required skills, a notice will be send to all educators who registered for the system, and collect their feedbacks. The feedbacks will then be analyzed to update the course map in the database.

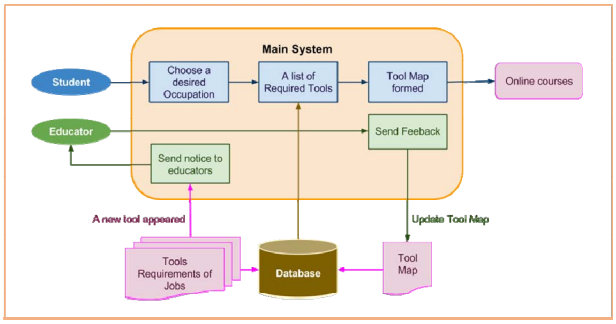

Figure 1 System Structure

A. Data Collection

In Taiwan, there are some popular job hunting web sites, which are also referred as an employment agency. An employment agency is an organization which matches employers to employees, where organizations posts job openings on the job hunting web sites with a certain amount of charge. Regular users are allowed to register with their identification number, then, write an own curriculum vitae in a personal page. The form of both data, from the organizations and users, are uploaded uniformly. Therefore, there is a standard input for the required skills. An example of the metadata of part of the "Computer/Internet" section is shown on Figure 2.

Computer/Internet								
Operating System	Software Development	Programming	Database Design	Server	Web Technology	Internet Technology	Office Application	Typesetting
AIX	ABAQUS	A+	Adabas	AS/400	ActiveX	AdvanceLink	Adobe Acrobat	FrameMaker
Apple	DDK	ActionScript	ADO	BizTalk	Apache SOAP	Asynch	Communicator	Adobe InDesign
DOS	MCU	ADA	ANSI SQL	CC Mail	Cold Fusion	Banyan	Excel	Pagemaker
FreeBSD	OOAD	AJAX	Brio	CICS	DHTML	Banyan Vines	ForeHelp	QuarkXPress
HP-UX	OOP	ASP	Capacity Plannin	Citrix	Dreamweaver	Bay	Ghost	
IDMS	Oracle Forms	ASP.NET	CMMS	ClearCase	EJB	BGP	Internet Explorer	
Linux	PVCS	ATL	Cognos	ClearQuest	Electronic Payme	Bridges	Lotus 123	
Mac OS	SDLC	C	Data Guard	Domino	Fireworks	BroadVision	Netscape Communication	
Mac/Macintosh	Servlets	C#	Data Modeling	FileNet	FrontPage	Checkpoint	OneNote	
Mainframe	STL	C++	Database Admini	Focus	GoLive	Cisco	Oracle Financials	
Microsoft SmartP	Systems Adminis	C++.Net	Database Manag	Hyperion (Brio)	HTML	DHCP	Outlook	
NDS/Novell Direc	Systems Analysis	CGI	DataStage	Microsoft Exchan	J2EE	DNS	PowerPoint	
Novell	Systems Analyst	Clipper	DB2	Microsoft ShareP	J2ME	e-commerce	Project	
OS X	UML	COBOL	Dbase	MQSeries	J2SE	EDI	Publisher	
OS/2	VxWorks	COM/DCOM	Endevor	Silverstream	JavaScript	Ethernet	Visio	
OS/390		COOL:Gen	ERwin	Tomcat	NetObjects Fusio	Firewall	Word	
OS/400		CORBA	Essbase	Vmware	RoboHelp	Frame Relay	Wordperfect	
Palm OS		Delphi	ETL	VSAM	SGML	FTP	WPS	

Figure 2 An example of the metadata of part of the "Computer/Internet" section

104 Job Hunting Site is one of the most popular job hunting site in Taiwan. The site is well organized and has uniform structured job description page. The searching results under computer science section will be crawled. Since the site is structured format, we can use web crawler to retrieve all the information of the job description. Using the html file of the searching results under computer science section, and find the titles of the researched results, we used python and beautiful soup to identify different attribute in the html file, and we can see the high lightered title is under the html tag "title", by using the similar method, we need to write the code that fit this specific searching results page. Once we get the title and the linked web link for the search results, then we will be able to look at the next level of the searched results, which is the actual content of the job requirement of the specific jobs. The metadata in the job requirements are classified as follows: Job Title, Content, Category, Experience, Education, Major, Language, Tools, Skills and Others. An example of a detailed job requirement page is shown on Figure 3. Here we know that the selections of tools are actually chosen from a fixed tick-box of all tools, therefore, we are able to exact the required tool list data from the html data. The extracted tools will be formatted in English. In the end, each job description will have a set of required skills and tools for the data clustering.

Figure 3 An example of a detailed job description page

B. *Sorting of Required Skills*

After the data has been collected, the grouping method for the required skills will be one of the major contributions. A certain job title can be posted on the job hunting web site, but this specific job title may be from different organizations, and for example, if we find three different organizations posting a recruit for software engineer, it is unlikely that the required skills for these three organizations will be the same, since different organizations may use different tools to develop. A possible method that can be used in this situation is called Jaccard Similarity. The Jaccard Similarity Coefficient is a parameter used to compare characteristic similarity between sets of information. Similarity measurement of Jaccard's between two example sets is a quotient of sharing characteristic number divided by all characteristic number. Therefore, the Jaccard's Similarity of all the skills sets collected are calculated.

	Java	JavaScript	C++	MS SQL	Linux	C#
Java		0.20401	0.16040	0.15017	0.16727	0.09589
JavaScript			0.05194	0.16727	0.06643	0.13433
C++				0.04878	0.19247	0.07984
MS SQL					0.06015	0.28899
Linux						0.05578
C#						

Figure 4 An example table of the Jaccard's Similarity calculated from the data collected.

Even though we have the Jaccard's Similarity calculated, it is difficult to define the vectors for algorithms like K-means. In the similar researched, most of the vectors derived from TF-IDF, which is called term frequency–inverse document frequency, a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval, text mining, and user modeling. Most of the researches find the related data from the job description using TF-IDF. But since we already have a better and organized job hunting site in Taiwan, and the skills or tools are fixed, we will have the direct and correct skills and tools required, without trying to extract useful data from the job description text.

C. *Clustering of Results*

Once we calculate the complete Jaccard's Similarity of the full tool matrix, we can start with the clustering of data. Here, we cannot use commonly seen methods of clustering such as K-means and so forth, because we cannot clearly define a way to measure the vector in this tool matrix. Then, we have tried Hierarchical Clustering on the tool matrix, the results is fine, but with a major problem where the same tool will not appear in another cluster. This clustering method will then not be suitable for our situation, because one skill can be needed as much important together two set of different skills. So, this clustering method will kill the opportunities for one tool to be classified into another tool set. In other words, another method has to be developed in order to suit this set of data. The method are explained below.

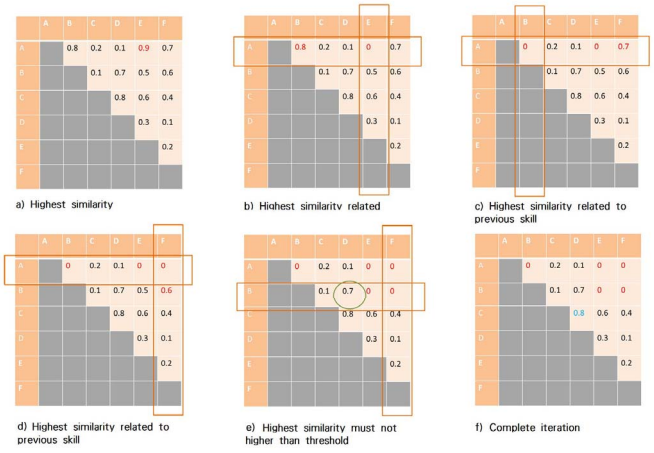

Figure 5 Diagram of the clustering algorithm

The steps explanation of the clustering algorithm used is showed on Figure 5. This example consists of the Jaccard's similarity of skills A to F. In order to find the first and the strongest cluster of skills, the pair of skills with the highest similarity is selected in diagram 5a, which is skills A and E with a value of 0.9. During the clustering, a threshold is set and modified to find the best results, in this case, the threshold is set to be 0.5. In each

iterations, the value selected must not be lower than the threshold. The second step would be to find the highest similarity related to both skill A and E, which is 0.8, the similarity between A and B, showed on diagram 5b. The same idea continues in the next steps, which is showed on diagram 5c and 5d. In the next step, showed in diagram 5e, please take note that, the value 0.7, circled in green, is not selected, otherwise the iteration will continue until all the values are selected, this is rule that when we selected the highest similarity value, we do not select the value lower than the threshold. At the same time, we always select the value that lower than the previous selected value. Diagram 5f shows the end of one complete iteration, this results in the first cluster found. Therefore, the first cluster consists of skills A, B, E and F. The next cluster will start with the value 0.8, which is the similarity between skills C and D marked in blue. By using this algorithm we can solve the problem happened when using Hierarchical Clustering, which is when the same skill do not appear in other clusters, this result will not be suitable in the real situation. When the above example is clustered, the results would be cluster 1{A, B, E, F}; cluster 2 {C, B, D} and cluster 3{C, E, F}. In this case, we will have a result that the same skill is possible to appear in two or more clusters.

Clustering Algorithm

```
while max Sim value > threshold
    for eachRow in range of All skills
        for eachCol in range of All skills
            find max Sim value
            record skills
    if max Sim value > threshold
        find next max Sim value
        add skill to list
end while
```

Figure 6 Pseudo code of the Clustering Algorithm

D. *Creating the Course Map*

Another major contribution is to form a course map from the results of the groups formed, and help the students by recommending the order of the courses to study these required skills. The "Computer Science Curricula 2013, Curriculum Guidelines for Undergraduate Degree Programs in Computer Science, December 20, 2013, The Joint Task Force on Computing Curricula Association for Computing Machinery (ACM) IEEE Computer Society" will studied to find out the inter relations between the courses, and a standard course map will be constructed according to this document. However, it is necessary for the course map to be updated if a new skill is required by the industry. A possible method to solve this problem is to implement a system that allows the educators to choose the possible parent nodes on the course map for the new skills. Therefore, whenever a new metadata of a required skill appears in the results of Sorting of Required Skills, a notice will be sent to the educators who are also part of this project or who had registered in this system, to encourage them to choose some possible parent courses for a specific skill based on their own educating experiences. A final result will then be analyzed from all feedbacks of the educators, and the new skill will be updated on the course map according to the final result.

IV. EXPERIMENT RESULT

First, the data are collected from www.104.com.tw, which is the popular job hunting site in Taiwan, which is collected every two weeks. Each set of the data from the Internet Technology section will have around 1500 of job recruitment. Each job recruitment will have a set of skills, these skills are then clustered using the clustering algorithm mentioned in section III C. Figure 7 shows a portion of the clustering results. The value of threshold is adjusted while observing from the results.

Figure 7 Sets of clusters of the Clustering Results

The results will be displayed on a web site. The skills are presented in the form of a course map, this is to give them a guide of the order to learn the courses, although the skills are not always related from the clustering results. Figure 8 shows the web presentation of the clustering results. On the left hand side, a scroll list will allow the selections of different type of clusters. When the type is selected, the course guide map will show on the right hand side. The course guide map showed is derived from the whole map which is created in section III D. Figure 8 shows a small portion of the skills from the whole map. When each skill node is clicked, it will direct to another page, where the searched results of the skill on the skill node selected, from the MOOCs courses will be showed. So, it is more convenient for students to go straight to have a glance at the actual course materials. When a new skill appears, the modification system which

will send notice to the users, are still in developing stage, and will soon be completed in the future.

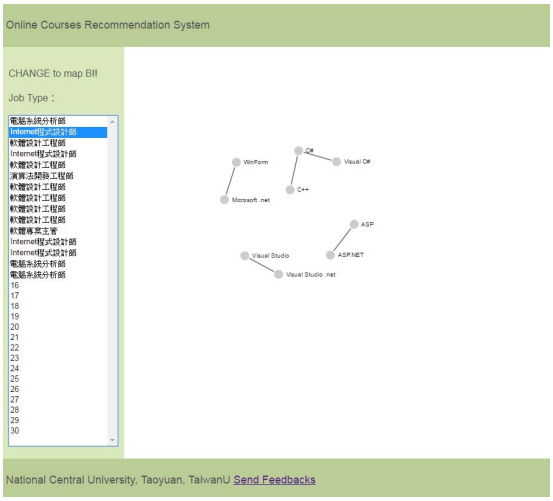

Figure 8 Web Presentation of the Clustering Results

COCLUSION

This system hopes to solve the problem known as the gap between industry and education, which the students, who had graduated from universities, do not always have the skills that the industries needed. In most cases, students will only have some skills or knowledge about some tools that is listed from the requirements of the industries. We encourage the students to self-study the courses according to the course map formed from their desired occupation, and therefore increase the chances that they will get the job offer. In order to show evidence that the system is helpful, and the clustering results are reliable, a feedback questionnaire will be collected, once the collected results are sufficient, a graph then can be shown to provide a stronger proof.

REFERENCES

[1] A. A. Bakar and C. Y. Ting, "Soft skills recommendation systems for IT jobs: A Bayesian network approach" on Conference on Data Mining and Optimization (DMO), Selangor Malaysia, 28-29 June 2011

[2] G. Poornalatha and Prakash S. Raghavendra, "Web user Session Clustering Using Modified K-Means algorithm", A. Abraham et al. (Eds.): ACC 2011, Part II, CCIS 191, pp. 243–252, Springer-Verlag Berlin Heidelberg 2011

[3] J. Singthongchai and S. Niwattanakul, "A Method for Measuring Keywords Similarity by Applying Jaccard's, N-Gram and Vector Space", Lecture Notes on Information Theory Vol. 1, No. 4, December 2013

[4] G. Karypis, "Evaluation of Item-Based Top-N Recommendation Algorithm", NSF CCR-9972519, EIA-9986042, ACI-9982274, by Army Research OÆce contract DA/DAAG55-98-1-0441, by the DOE ASCI program, and by Army High Performance Computing Research Center contract number DAAH04-95-C-0008, 2011

[5] Y. Wang, X. Zhang, L. Nan, and D. Wang "Occupation recommendation based on student achievement mining in vocational skill training", 11th International Conference on Fuzzy Systems and Knowledge Discovery, 2014

[6] G. Linden, B. Smith, and J. York, "Amazon.com recommendations- item-to-item collaborative filtering", IEEE Internet Computing, January • February, 2003

[7] Juan J. Cameron, Carson Kai-Sang Leung and Syed K. Tanbeer, "Finding Strong Groups of Friends among Friends in Social Networks", on ninth IEEE International Conference on Dependable, Autonomic and Secure Computing, 2011.

[8] C. Hauff and G. Gousios,"Matching GitHub developer profiles to job" on 12th Working Conference on Mining Software Repositories, 2015

[9] Y. Zhang, C. Yang and Z. Niu, "A Research of Job Recommendation System Based on Collaborative Filtering" on Seventh International Symposium on Computational Intelligence and Design, 2014

[10] J. Wang, Z. Liu and H. Zhao, "Group Recommendation Using Topic Identification in Social Networks" on Sixth International Conference on Intelligent Human-Machine Systems and Cybernetics, 2014

[11] M. S. Chen, J. Han and Philip S. Yu, "Data Mining: An Overview from a Database Perspective" on IEEE Transactions on Knowledge and Data Engineering, Vol 8, No 6, December, 1996

[12] S. Ye, J. Lang and F. Wu, "Crawling Online Social Graphs" on 12th International Asia-Pacific Web Conference, 2010

[13] H. Sun, T. Wu, M. Yan and Y. Wu, "A new item clustering-based collarative filtering approach" on Ninth Web Information Systems and Applications Conference, 2012

[14] Y. Li, J. Zhang and H. Dan, "Text Clustering Based on Domain Ontology and Latent Semantic Analysis", on International Conference on Asian Language Processing, 2010

[15] Nguyen Viet Anh, Nguyen Viet Ha And Ho Si Dam, "Constructing A Bayesian Belief Network to Generate Learning Path In Adaptive Hypermedia System" , 2008

[16] J. Davies and M. Graff, "Performance in e-learning: online participation and student grades", British Journal of Educational Technology Vol 36 No 4 2005 657–663, 2005