# Enhanced Clustering-based MOOC Recommendations using LinkedIn Profiles (MR-LI)

Fatimah Alruwaili, Dimah Alahmadi
Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah
Saudi Arabia

***Abstract*—With the rapid development of massive open online courses (MOOCs), the interest of learners in MOOCs has increased significantly. MOOC platforms offer thousands of varied courses with many options. These options make it difficult for learners to choose courses that suit their needs and compatible with their interests. So, they become exposed to many courses on all topics. Therefore, there is an urgent need for personalized recommendation systems that assist learners in filtering courses according to their interests. Therefore, in this research, we target learners on the professional platform, LinkedIn, to be the basis for user modeling; the number of extracted profiles equals 5,039. Then, skill-based clustering algorithms were applied to LinkedIn users. Subsequently, we applied the similarity measurement between the vector features of the resulting clusters and the extracted course vectors. In the experiment result, four clusters were provided with the top-N course recommendations. Ultimately, the proposed approach was evaluated, and the F1-score of the approach was .81.**

***Keywords—MOOCs; recommendation systems; content-based; clustering; term frequency-inverse document frequency (TF-IDF); LinkedIn***

## I. INTRODUCTION

Currently, the world has significantly developed regarding the services provided to learners over the internet. These services have expanded to include courses, academic qualifications, and science lessons and have become known as e-learning. After the emergence of e-learning, many students worldwide have participated in online courses in virtual classes [1]. Recently, E-learning has gained colossal attention since the emergence of Massive Open Online Courses (MOOCs), which attract numerous learners to engage[2]. MOOCs are an online platform that provides services for learners of different ages and academic levels worldwide in different geographic locations.  It serves the community's largest possible segment and has more than 100 million students [3]. Some MOOCs are offered as free educational courses for learners in various fields in many languages. It is also characterized by its flexibility to accept students and facilitate their access to available educational content. MOOCs offer different styles to deliver courses where they can access course content in text or video lectures. They can take advantage of extending course content by using discussion boards or blogs [4][5]. MOOC evolutions are considered the most popular platforms that have evolved to include all countries globally, with many providers such as Coursera, Udacity, Udemy, and Edx [6].

However, due to the large number of provided courses, the learners may face difficulties obtaining the desired content. In terms of choosing the right course, the number of users who make a wrong decision exceeded 90% [7]; thus, a meaningful recommendations engine has become critical for MOOC users. Based on these facts, the importance of personalized recommendations for users in education or other fields should be mentioned. The recommendations provided to learners have great significance and may be one of the most critical factors motivating them to expand their learning experience with various courses offered to them. MOOC recommendations have importance for both sides, learners, and MOOC providers, as learners face difficulty reaching the appropriate content. At the same time, MOOC providers also face problems represented in suggesting the proper course. There are different recommendation techniques; some recommendations can be achieved using collaborative filtering methods, which provide the learner with recommendations similar to the courses that their peers joined in the platform. Other systems have relied on user modeling by analyzing their search history in the platform or analyzing their profile in the MOOC platform [8]. Recent studies have confirmed that the most effective methods in the recommendations are the ones that rely on the analysis of social networks because it is closer to match the taste of users. Recommendation systems used social network data to give the user more customized recommendations based on each user's personal information [9]. This research has relied on the utilization of social network content to customize recommendations. Specifically, LinkedIn was chosen to be the primary source for this research's dataset for many reasons. First, it is one of the best professional social networks where users express their education, academic experience, skills, and educational interests. The proposed approach in this paper analyzes users' profiles on LinkedIn and then provides course recommendations for the most appropriate course of these profiles.

The paper will be organized as follows: Section 2 will discuss the related works to MOOC recommendation systems, especially content-based systems. Then, Section 3 will discuss our proposed approach to courses recommendation based on LinkedIn data, starting with data collection, description, cleaning, the clustering process for LinkedIn profiles, and the recommendation process. Then, the evaluation process of the proposed approach will be discussed in Section 4. After that, we will discuss the results in Section 5. Finally, in Section 6, we will conclude the work and present the future directions for this research.

## II. Literature Review

Due to the knowledge and information rapid explosion worldwide, there is an urgent need to improve the efficiency of the learning process. Therefore, MOOC platforms became more popularized to fulfill this need equipped with recommendation systems.

Recent studies have also confirmed the effectiveness of integrating information derived from social networks with recommendation systems in terms of accuracy. The additional information about the user increases the understanding of the user's behavior and preferences. Thus, the user can be better understood and modeled, reflecting positively on the accuracy of the recommendation [8][10]. Data mining has helped researchers develop recommendation systems (RSs) to provide users with suggestions related to specific items or content to achieve personalization [11]. Many studies were initiated to assist in recommend courses for learners. For example, Dumitru Radoiu [7] addressed the user attributes, user behaviors, and item attributes in MOOC platforms such as ‘user profile’ (user attribute), ‘user history’ (user behavior), and ‘course description’ (item attribute) to provide learners with suitable course recommendations to improve their completion rate.

Other ways to use recommendation systems in e-learning, as in the study by Kardan et al. [2], which analyzed social networks to lead learners to match relevant information in MOOC platforms. Additionally, in terms of the efforts to solve the cold start problem, a study by Kumar Abhinav et al. [12] presented a framework of hybrid recommendation systems. Many predictive models have been used to contribute to providing efficient course recommendations for learners. At the same time, other studies such as Xiao Li et al. [13] have applied the user preferences and behavior inside the demographics data to develop accurate recommendations to serve their needs. The definitive study by Alzahrani & Maccawy [14] proposed a hybrid model for MOOC search based on personalization known as the MOOC Recommender Search Engine (MRSE) to access relevant courses easily. In a new study in 2021, Khalid et al. [15] proposed an algorithm based on ratings. The system implements a new algorithm characterized by flexibility and scalability; what is more, it is more accurate than previous algorithms. Its results were also compared with the Collaborative Filtering and Clustering algorithms, and they showed great superiority in the accuracy and classification metrics.

Furthermore, in content-based recommendations algorithms, the item's content is used to provide recommendations; it includes the information to describes the items [11]. Many studies have adopted this technique in building MOOC course recommendation models for learners. For example, a study by Piao & Breslin [16] presented a system that gives learners personalized recommendations by taking advantage of their LinkedIn pages' data. This system ranks the courses obtained from Coursera (using google custom search engine GCSE) according to its similarity with the user’s profiles. Jing & Tang [17] developed a new algorithm called Hybrid Content-Aware Course Recommendation (HCACR); they employ collaborative filtering to develop a course recommendation algorithm that combines user interests and demographics as well as analyzes pre-course requirements. They tested the proposed algorithm on the "XuetangX" platform [18], a Chinese courses platform; their algorithm proved its effectiveness, as it achieved a high click rate on the recommended courses. Another study by Huang & Lu [19] presents a content-based MOOC model for intelligent education, contributing to user profiling development. They have used user interest analysis on the MOOC page to create a user profile and provide recommendations that match the user's activity log. In another study by Zhang et al. [20], the authors developed a recommendation model based on content analysis for learners and educational courses (named MOOCRC), which relies on deep belief networks (DBNs). This graphical model combines probability and statistics with machine learning and neural networks in MOOC environments. Their proposed model achieved higher accuracy and coverage rate than the traditional recommendation systems. Using a context-aware factorization machine algorithm, Chanaa & El Faddouli [21] designed a new recommendation approach for a MOOC platform in order to provide further recommendations that aligned with each learner using predictions about user behavior; this was studied by analyzing user interactions in the MOOC platforms, including rating, feedback, and likes.

In specific applications, many researchers used LinkedIn in recommending MOOC courses. LinkedIn offers the opportunity to obtain the user’s profile in order to analyze the user's educational taste. Users' profiles contain reliable information about learners’ scientific backgrounds and research fields, which is considered the largest professional social network on the internet [22]. Besides the valuable information existing in user profiles, such as the educational degree and work experience [23]. In a similar study of Dai et al. [24], the data was collected and analyzed from the professional profiles on LinkedIn. The authors used the natural language processing techniques (NLP) to study the users' behavior on Online Social Networks (OSN) to provide recommendations that improve their decision-making process. Also, Dai et al. [25] used the available personal data on LinkedIn pages to provide customized recommendations to learners based on their preferences. However, these preferences were built by focusing on the job market to make the recommendations more relevant to the job market's needs. Another study for Pourheidari et al. [9] used data taken from two well-known social networking sites, LinkedIn and Twitter, to provide users with recommendations that essentially correspond to their information written on LinkedIn and Twitter. This study proved to be highly effective in recommendation systems. The last research was by Kumalasari and Susanto [26], who collected data from professional profiles on IT professionals from LinkedIn to be used as a reference for the skills presented later as a recommendation for students (job seekers).

### A. Research Gap

The learners had difficulty in obtaining the appropriate training courses. There have been many studies to solve this problem, and one of the most effective ways is to provide personalized recommendations for learners. Therefore, this research will present a personalized recommendations approach for learners based on their skills. According to the literature, the studies that used the content-based recommendation had a good performance. Other studies that used social networks in personalization also achieved recommendations closer to the learner's needs. Therefore, in this research, these two features will be combined to personalize courses for learners better.

The content-based recommendation will be used as well as the applying of clustering algorithms besides utilizing the TF-IDF technique. Recommendations will also be personalized based on social networks. Therefore, in this research, the professional social network "LinkedIn" will be relied upon, as it is the most formal social network; besides focusing on the user's skills present on LinkedIn. In addition, the studies that highlight users' skills in LinkedIn are limited, so the skills here will be used as a guide to the course's recommendation process since it was closest to describing the user's interest. So, the contribution of this research will be to apply clustering algorithms to LinkedIn users to provide personalized course recommendations to learners based on their profiles, especially their skills present on LinkedIn.

## III. MOOC Recommendation based on LinkedIn Profiles KR-LI Approach

### A. Proposed Approach

This section presents our approach "MOOCs Recommendation based on LinkedIn MR-LI" to recommend MOOC courses to learners based on their LinkedIn profiles. This approach aims to identify the learner's interests through the mentioned skills in him/her profile that explicitly expresses the scientific field in which he/she is interested. Fig. 1 illustrates the general framework for the proposed approach.

First, the approach extracts data from the LinkedIn and Coursera websites. The crawling process is performed on LinkedIn profiles to scrape the entire information from each profile, and then store the scrapped data into a user's dataset. As well as for Coursera, then it is stored in a separate dataset for the courses. Second, the approach begins with the cleaning and preprocessing of datasets. Third: LinkedIn users are clustered into clusters based on the similarities between the users. At this step, the skills field on LinkedIn is taken as a feature for clustering on its basis. Fourth: begins with the feature extraction for datasets (LinkedIn and Coursera) and the calculation of weights for feature vector construction using term frequency-inverse document frequency TF-IDF. Fifth, the similarity between the learners and the courses is measured using Cosine Similarity. Finally, the approach provides recommendations for learners' clusters with ten courses that are most similar to these clusters. The following section will discuss these steps in detail.

### B. Dataset Collection

*1) LinkedIn dataset:* In order to validate MR-LI, we used LinkedIn as the primary resource for our learners' dataset. This is because LinkedIn is one of the most popular social network sites in which people express their interests and educational backgrounds more formally and professionally, as it is specific to employment and education development, so it is the best environment to obtain accurate data for the recommendation process [9].

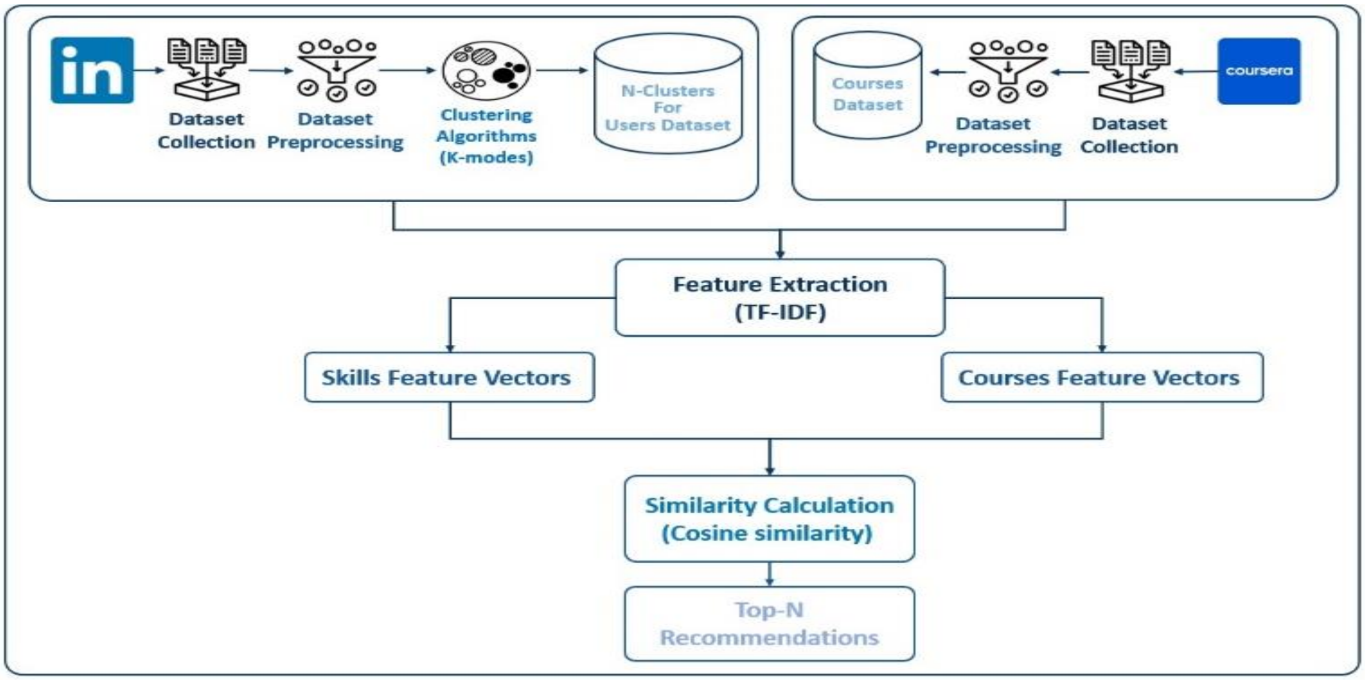
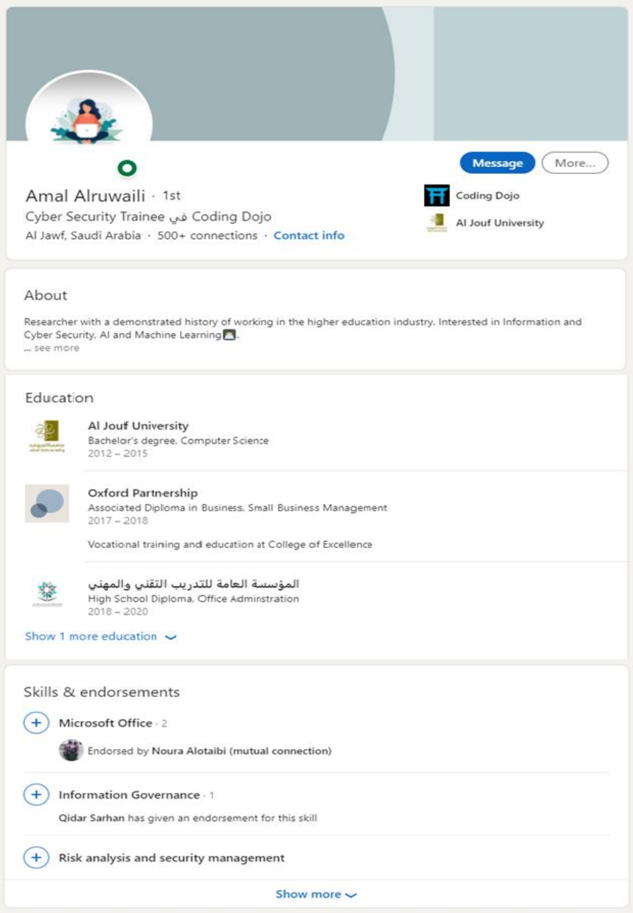


Fig. 1. The Framework for the Proposed Approach MR-LI.

We used Python as the programming language to deal with the LinkedIn API. The BeautifulSoup and Selenium libraries were used to access users' profiles [24] and then store the data in JSON and CSV format. Using LinkedIn API, we can access public data for users such as name, current job, past jobs, degrees, brief description, skills, interests, languages, etc. Fig. 2 provides an example of a public profile on LinkedIn. There are two profile types on LinkedIn (public & private), and we can access the public profile only. In order to extract our dataset, we identified the subscribers in common companies and universities in Saudi Arabia to reach the actual active users in Saudi Arabia. After the data scrapping process, the number of users reached more than 20,000. Table I shows the most important information on users' LinkedIn profiles.



Message More...

Amal Alruwaili · 1st
Cyber Security Trainee في Coding Dojo
Al Jawf, Saudi Arabia · 500+ connections · Contact info

Coding Dojo
Al Jouf University

About

Researcher with a demonstrated history of working in the higher education industry. Interested in Information and Cyber Security, AI and Machine Learning.
... see more

Education

Al Jouf University
Bachelor's degree, Computer Science
2012 – 2015

Oxford Partnership
Associated Diploma in Business, Small Business Management
2017 – 2018

Vocational training and education at College of Excellence

المؤسسة العامة للتدريب التقني والمهني
High School Diploma, Office Adminstration
2018 – 2020

Show 1 more education

Skills & endorsements

Microsoft Office · 2
Endorsed by Noura Alotaibi (mutual connection)

Information Governance · 1
Qidar Sarhan has given an endorsement for this skill

Risk analysis and security management

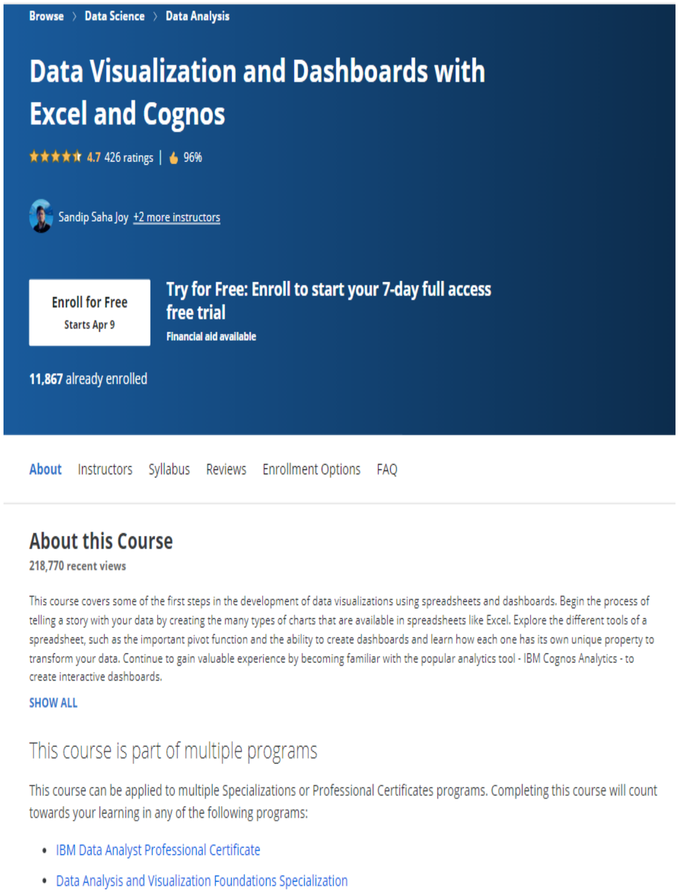
Show more

Fig. 2. Example of Full Profile on LinkedIn.

TABLE I. THE MOST IMPORTANT INFORMATION ON USERS' LINKEDIN PROFILE

| Field Name | Description |
|---|---|
| About | Brief introduction about the user. |
| Activity | Posts the user publishes and the posts he/she interacts with. |
| Experience | Practical user experiences. |
| Education | Educational qualifications obtained by the user. |
| Licenses & Certification | Professional licenses and certificates are obtained by the user. |
| Skills | Skills the user possesses. |

*2) Coursera dataset:* Conversely, to obtain data for the courses that will be recommended to learners, we have chosen the Coursera website [27], one of the largest global platforms that offer courses in various technology fields and others. Coursera provides details about the courses on each course page, as in Fig. 3. Therefore, it is considered an excellent platform in terms of the details available about the offered courses. The API with BeautifulSoup and Selenium libraries on Python also have been used to scrape 12173 courses in JSON and CSV format. Table II shows the most important information on the course page on the Coursera website, such as the course title, description in "about this course", instructors, etc.



Browse › Data Science › Data Analysis

Data Visualization and Dashboards with Excel and Cognos

★★★★★ 4.7 426 ratings | 👍 96%

Sandip Saha Joy +2 more instructors

Enroll for Free
Starts Apr 9

Try for Free: Enroll to start your 7-day full access free trial
Financial aid available

11,867 already enrolled

About Instructors Syllabus Reviews Enrollment Options FAQ

About this Course
218,770 recent views

This course covers some of the first steps in the development of data visualizations using spreadsheets and dashboards. Begin the process of telling a story with your data by creating the many types of charts that are available in spreadsheets like Excel. Explore the different tools of a spreadsheet, such as the important pivot function and the ability to create dashboards and learn how each one has its own unique property to transform your data. Continue to gain valuable experience by becoming familiar with the popular analytics tool - IBM Cognos Analytics - to create interactive dashboards.

SHOW ALL

This course is part of multiple programs

This course can be applied to multiple Specializations or Professional Certificates programs. Completing this course will count towards your learning in any of the following programs:

- IBM Data Analyst Professional Certificate
- Data Analysis and Visualization Foundations Specialization

Fig. 3. Example of Course on Coursera.

TABLE II. THE MOST IMPORTANT INFORMATION ON THE COURSE PAGE

| Field Name | Description |
|---|---|
| Course Title | Title of the course. |
| About | General description of the course and its contents. |
| Instructors | Details about the instructors presenting the course. |
| Syllabus | Details about the course content. |
| Review | Learners' rating and feedback about the course. |
| Enrolment Options | Options for attendance and payment methods. |

### C. Dataset Cleaning and Preprocessing

*1) LinkedIn dataset:* The initial total of data extracted was over 20,000 files. In order to obtain satisfactory results, files that do not contain primary data for the recommendation process include: profileurl, firstname, lastname, schooldegree, schooldegreespec, schooldegree2, schooldegreespec2, allskills, skill1, skill2, skill3, skill4, skill5, and skill6 are excluded. Also, the number of profiles written in Arabic was scarce due to the reliance of the majority of users on writing their profiles in English. Because the small dataset number did not yield satisfactory results, we had to exclude the Arabic profiles. Therefore, emphasis was placed on the English profiles only; the profiles written in other languages were excluded.

In addition, we performed some preprocessing on the data: first, transforming the text to lowercase. Second, removing the punctuations, stop words, and URLs. Third, excluding meaningless rows with long descriptions from skills or not writing them in text. Fourth, excluding profiles that contain less than six skills. Finally, separating the skills using "|". The data size after cleaning amounted to 5,039 files.

LinkedIn gives its users complete freedom to express themselves, their skills, academic qualifications, and experiences [24]. Therefore, there is no specific way to write the skills. For example, we find that a person may express 'Leadership' in 'Team Leadership', 'Team Management', or 'TeamLeadership'. In this way, the writing style can affect the distribution of users in clusters. Therefore, we normalized the skills as shown in Table III. After that, we performed the lemmatization process on the dataset to avoid data duplication. By the end of this process, we found that the common skills were "Microsoft Office", "Project Management", "Teamwork", and "Leadership".

TABLE III. THE NORMALIZATION FOR LINKEDIN DATASET

| Skill | Keywords |
|---|---|
| 'Programming' | 'Program', 'programming' |
| 'Project Management' | 'Project', 'Project management', 'PMP', 'End-to-End Project Management' |
| 'Teamwork' | 'Team', 'Team work', 'Team Work', 'team work' |
| 'Leadership' | 'Team Leadership', 'Team Management' |
| 'Time Management' | 'Time', 'Time Management', 'time management', and 'TimeManagement'. |
| 'Accounting' | 'Accounts', 'Accountant', and 'Accounting'. |
| 'Microsoft Office' | 'Office', 'Microsoft', and 'microsoft'. |
| 'Presentation Skills' | 'presentation', 'presentation skills'. |
| 'Strategic Planning' | 'Strategy', 'Strategic', |
| 'Data Analysis' | 'Data', 'Data Analytics', |
| 'Financial Analysis' | 'finance', 'Finance', 'Financial' |
| 'Web Development' | 'Web', 'webdenepment', 'webdeveloper'. |
| 'Business Development' | 'Business', 'Business developer', 'Business Skills'. |
| 'Quality Assurance' | 'Test', 'Tester', 'Quality', 'Assurance of quality'. |
| 'Object-Oriented Programming (OOP)' | 'OOP', 'Object-Oriented Programming' |

*2) Coursera dataset:* The scraped data contains the following information: CourseId, Description, CourseTitle, DurationInSeconds, ReleaseDate, AssessmentStatus, IsCourseRetired. However, the primary data for each course is the CourseId, CourseTitle, and Description columns, so courses that do not contain this information have been excluded. Also, the "IsCourseRetired" column represents the course's state in real-time is available or not. So, the courses with value = "No" in this column were excluded from the recommendation process to avoid making recommendations to learners with unavailable courses. The data have been cleaned of stop words, symbols, punctuation marks, and all characters except numbers and letters. Also, terms like "ll" used in "Description" texts, such as 'we'll' and 'you'll', were also removed along with '-' (hyphens) from the CourseId. The data size after the cleaning equaled 3,471 courses.

### D. Clustering LinkedIn Profiles

The purpose of this section is to categorize users based on their skills. Nevertheless, due to the freedom granted to users by LinkedIn, they can express their skills in various names without using pre-defined labeling. So, LinkedIn profile fields do not follow a specific standard, such as the UNESCO used to classify the users' skills [28]. In this sense, the classification algorithms become very difficult. So, the solution here is clustering algorithms, as it clusters the users according to how similar they are to each other. Considering the size of the obtained dataset, we decide to apply the K-modes clustering method due to its efficiency and effectiveness in the used size of the dataset. Before using the K-modes algorithm, we must determine the number of clusters "k" since it is a sensitive parameter for this clustering process [24]. We applied the Elbow method to guide the choice of the "k" parameter [29].

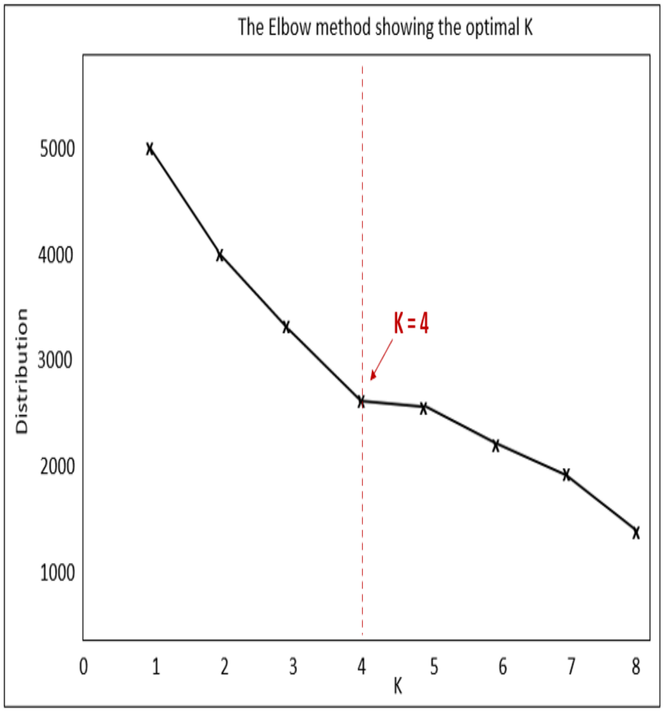


Fig. 4. The Result of the Elbow Method.

As shown in Fig. 4, the elbow method computes the squared distance's total for each cluster. We assign different k-values from 0 to 49; by analyzing the generated graph, it is clear that the k equals the breakpoint, which is the elbow point. In this case, according to the graph, the optimal k will be 4.

The K-Mode algorithm was applied to 5039 profiles. Table IV shows the number of profiles in each one of the four clusters.

We notice that more than 75% of the profiles are located in the first cluster. Therefore, a lemmatization process was performed on the dataset to reduce the similarity between profiles. Table V represents the distribution of the profiles on the four clusters after the lemmatization process.

We created word clouds corresponding to each cluster to clarify the distribution of the profiles in the four clusters. In the beginning, we notice the repetition of some skills in all clusters, such as "Microsoft Office", "Teamwork", and "Time management", but at different rates from one clause to another. Table VI shows the commonly used skills in each cluster.

For the first cluster, as shown in Fig. 5, it is clear that the most common skills among users are combined, and this explains why it is the largest cluster among the four clusters, as it contains "business," "planning," "problem-solving," and "communication" as the most common skills in this cluster. As for the second cluster in Fig. 6, it is clear that the most common skills are "human resources", "development", "recruit", "team management", and "social media", it can be described as it combines social, employment, and communication skills in general. As for the third cluster in Fig. 7., it is widely noted that it combines skills that indicate interest in the field of cybersecurity such as "defense", "protection", "threat", "awareness" and "iam", which are terms widely used in the field of cybersecurity. Finally, the fourth cluster in Fig. 8 gathered skills that generally referred to project management and software engineering. Fig. 9 shows the skills that are most frequently used among the users of the four clusters.

TABLE IV. THE NUMBER OF PROFILES IN EACH CLUSTER

| Cluster ID | Number of profiles |
|---|---|
| Cluster 1 | 3,859 |
| Cluster 2 | 392 |
| Cluster 3 | 367 |
| Cluster 4 | 448 |

TABLE V. THE NUMBER OF PROFILES IN EACH CLUSTER AFTER THE LEMMATIZATION

| Cluster ID | Number of profiles |
|---|---|
| Cluster 1 | 3,175 |
| Cluster 2 | 463 |
| Cluster 3 | 952 |
| Cluster 4 | 448 |

TABLE VI. THE COMMONLY USED SKILLS IN EACH CLUSTER

| Cluster ID | Common Skills | |
|---|---|---|
| Cluster 1 | Microsoft<br>excel<br>server<br>test<br>software<br>database<br>leadership<br>creative<br>integration<br>analytical | warehouse<br>business<br>quality<br>databases<br>android<br>teamwork<br>communication<br>problemsolve<br>selfconfidence<br>*management |
| Cluster 2 | teamwork<br>Leadership<br>hr<br>management<br>analysis<br>change<br>corporate<br>relationship<br>research<br>strategic | plan<br>recruit<br>coach<br>staff<br>project<br>data<br>analysis<br>quality<br>performance<br>documentation |
| Cluster 3 | cybersecurity<br>threat<br>identity<br>risk<br>ld<br>software<br>access<br>control<br>iso<br>disasterrecovery * | iam<br>vulnerability<br>assessment<br>delay<br>defense<br>data<br>protection<br>regulation<br>continuity<br>authentication |
| Cluster 4 | mysql<br>application<br>scalability<br>requirement<br>database<br>design<br>solution<br>information<br>technology<br>startup | agile<br>software<br>lifecycle<br>sdlc<br>intelligence<br>change<br>test<br>database<br>integration<br>scrum |

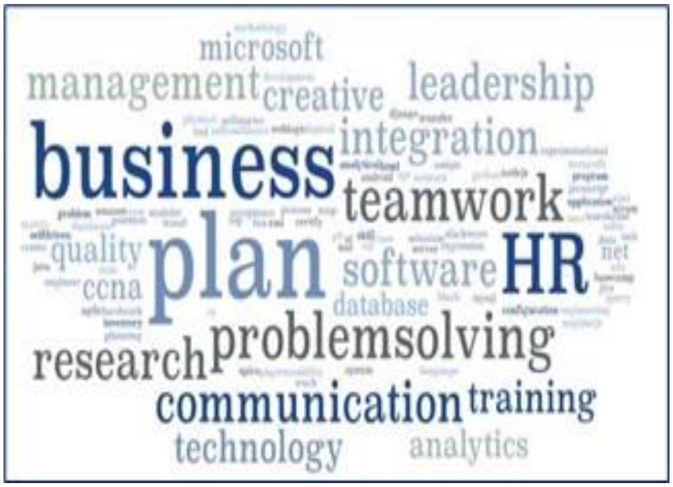


Fig. 5. Wordcloud of Cluster 1.

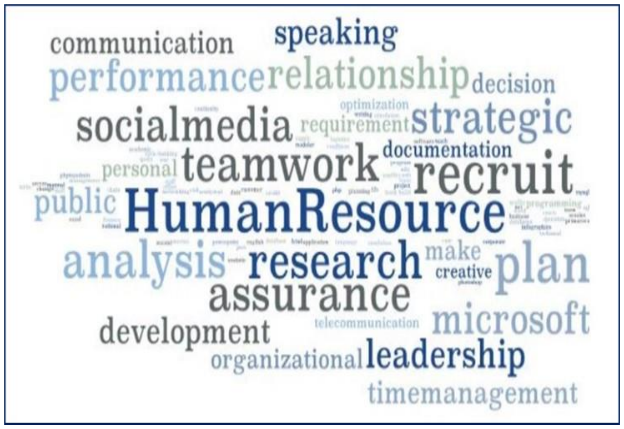


Fig. 6. Wordcloud of Cluster 2.

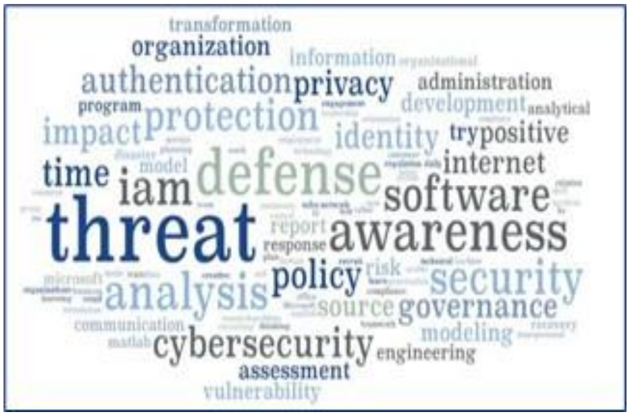


Fig. 7. Wordcloud of Cluster 3.

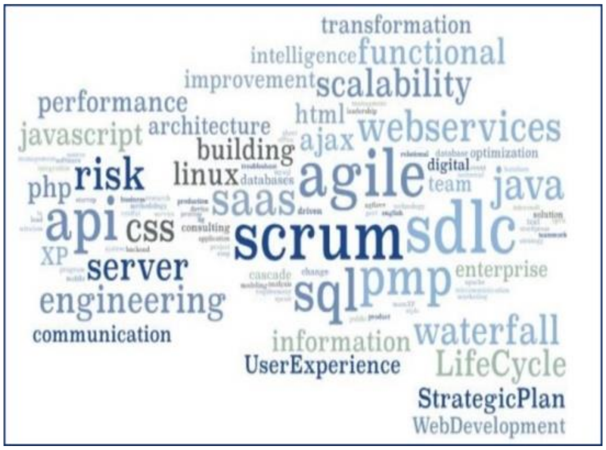


Fig. 8. Wordcloud of Cluster 4.

### E. Recommendation Process

This phase consists of two steps in which we aim to identify the recommended courses for the four user clusters.

*1) TF-IDF:* In order to extract the important features in the datasets, the term frequency-inverse document frequency (TF-IDF) was used, which proved to be effective in detecting important words for the dataset [30]. In the vector space model, TF-IDF is the commonly used weighting method in describing the documents [31].

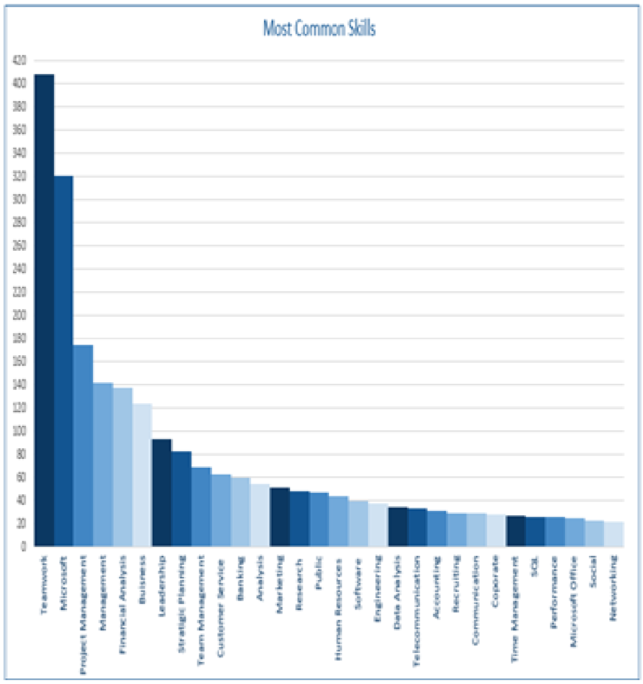


Fig. 9. The Most Commonly used Skills in the Four Clusters.

TF-IDF indicates the importance of the term for the whole document. It is related to the number of times the word appears in a document compared with its frequency in the document. Thus, Tf in TF-IDF weight measures the frequency of the terms in a document, while IDF measures the term importance in the document. The following equation illustrates the TF-IDF method.

$$tfidf_{i,j} = tf_{i,j} \times \log(\frac{N}{df_i}) \tag{1}$$

$tf_{i,j}$ = number of occurrences of i in j.

$df_i$ = number of documents containing i.

N = total number of documents.

Therefore, the importance of a word increases with the value of TF-IDF for that word. Thus, the higher the TF-IDF value for a specific skill in Cluster, the higher the value of this skill will be, and likewise for the courses data set, vice versa.

*2) Similarity measure:* In order to begin the recommendation process, the similarity between each of the four clusters should be measured with the courses in the course dataset. In this step, one of the most popular metrics used to measure similarity is the cosine similarity [32]. The formula is:

$$similarity = \cos(\theta) = \frac{A.B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2 \sum_{i=1}^{n} B_i^2}} \tag{2}$$

The similarity of each cluster is measured using the ten most similar courses in the course dataset. After that, these ten courses are presented as user recommendations in this cluster; this proceeds for all four clusters.

Subsequently, ten-course recommendations are presented for each cluster based on the results of cosine similarity, which are the top 10 similar courses for each cluster. Table VII represents the recommendations resulting from the four clusters.

TABLE VII. THE RESULTED RECOMMENDATIONS FOR EACH CLUSTER

| **Cluster ID** | **Recommendations** |
|---|---|
| Cluster 1 | Building Excel Online Automation with Office Scripts<br>Building Websites with HTML, CSS, and JavaScript: Getting Started<br>Build Your First Dashboard with GoodData<br>How Novices Learn to Program: What I've Learned Teaching in a Coding Bootcamp<br>Gin: A Website Application Framework for Go<br>Organizational Design: Going from Features to Experiences: Front 2019<br>Exploring Product Sales<br>Unlocking Unstructured: Leveraging Data Discovery<br>Creating and Using Track Mattes in After Effects<br>AWS Infrastructure with Python: Getting Started |
| Cluster 2 | PMP® Exam Prep â€“ Project Human Resource Management<br>Introduction to Presentation Design<br>PMPÂ® Exam Prep â€“ Project Communications Management<br>Computing, Communication, and Business Integration for CASP (CAS-002)<br>Managing Delivery of Your App via DevOps<br>Leveling up<br>Planning and Designing Microsoft Azure Networking Solutions<br>Website Performance<br>LinkedIn Fundamentals<br>Creating Animated Web and Social Media Banners in Photoshop and Flash |
| Cluster 3 | Building Your Cyber Security Vocabulary<br>Cyber Security Awareness: Malware Explained<br>Cyber Security While Traveling<br>Layer 2 Security for CCNA Security (210-260) IINS<br>The Issues of Identity and Access Management (IAM)<br>Incident Detection and Investigation with QRadar<br>CompTIA Security+ (2008 Objectives)<br>Cyber Security Awareness: Social Engineering<br>SSCPÂ®: Monitoring and Analysis & Risk, Response, and Recovery (2012 Objectives)<br>Preparing for the Google Cloud Professional Cloud Architect Exam |
| Cluster 4 | Driving Engineering Culture Change at Microsoft: An Experimental Journey<br>Scalable, Flexible, Modular, Preventative Architecture<br>Agile Estimation<br>Managing Work with Team Foundation Server 2012<br>Testing AngularJS from Scratch<br>Easily Estimate Projects Using Statistics and Excel<br>Secure Software Development<br>CISSPÂ® - Software Development Security<br>Windows 2000 Server Group Policy<br>Java: JSON Databinding with Jackson |

## IV. EVALUATING MR-LI APPROACH

At first, we used experts to carry out the recommendation process manually. We asked the experts to separately provide ten-course recommendations for each cluster by matching the skills in each cluster to the most appropriate courses based on the course description. Then, we provide them with the four clusters and skills in each cluster, with the weight of each skill besides the courses. The experts generated ten ordered recommendations for the four clusters. Next, to evaluate the performance of the proposed approach, we compared the results generated from the approach against the results from the experts. By comparing the results, we find that only two cases result from comparing the recommendations. The first is that the approach's recommendation matches the human recommendation, and we refer to this case as true (true is quantified by 1). The second case is the opposite: the experts give a recommendation that does not match the recommendation resulting from the approach; we refer to this case as false (false is quantified by 0).

In order to achieve this, we created two empty lists, "T" and "F", one for the values of the ones, the other for zeros, and to do this between results, the comparison is based on the equality of the match results in both ways; thus, if a recommendation from cluster1 for the approach as an example exists in cluster1, then the recommendation is correct. The value of '1' will be added to the "T" list, in the other case, it is a false recommendation, and '0 'will be added to the "F" list, and at the end, the accuracy is the number of correct recommendations divided on the total number of recommendations. The following equation illustrates this process.

$$Accuracy = \frac{(T)}{(T)+(F)} \tag{3}$$

As a result, the accuracy of the proposed approach was 0.675. In addition to accuracy, other measures are used to evaluate the statistical result, namely precision and recall. Precision calculates the percentage of the related documents with the selected documents illustrated in Equation 4. In contrast, the recall measures the percentage of the related documents compared to all related documents found in the selected documents and is shown in Equation 5. Therefore, after applying accuracy and recall, we can apply F-Measure as shown in Equation 6. Table VIII shows the results.

$$P = \frac{TP_i}{TP_i + FP_i} \tag{4}$$

$$R = \frac{TP_i}{TP_i + FN_i} \tag{5}$$

$$F = \frac{2 \times P \times R}{P + R} \tag{6}$$

As a result, the F1 score was .81, which means we have an excellent working approach for our recommendations if we consider that we are treating strings matching (Skills / Courses). And for the precision, we are trying to find how much trues exist in the positives, but we have the false positives = 0, so the precision was 1, to make it clear, FP = 0, because the human way never recommend a course which is false for a certain skill, then the false = 0 and that's why precision = 1. Finally, the recall was 0.68.

TABLE VIII. THE PERFORMANCE RESULTS

| | Precision | Recall | F1 Score |
|---|---|---|---|
| Cluster 1 | 1 | 0.50 | 0.67 |
| Cluster 2 | 1 | 0.70 | 0.82 |
| Cluster 3 | 1 | 0.80 | 0.89 |
| Cluster 4 | 1 | 0.70 | 0.82 |
| **MR-LI** | 1 | 0.68 | **0.81** |

## V. CONCLUSION

Social networks are extremely valuable in obtaining information that assists in modeling the user in a manner similar to reality. Therefore, LinkedIn, one of the largest professional social networks, provided customized course recommendations to users. These recommendations help users quickly reach the courses that suit their interests without requiring much search effort. This paper presents MR-LI as a course recommendation approach that relies on clustering algorithms to group the users according to their LinkedIn skills, resulting in four dataset clusters. Subsequently, feature vectors are extracted using TF-IDF for the user datasets and course datasets. The similarity was measured between each cluster's feature vectors and the courses using cosine similarity. The resulting ten recommendations were presented for each cluster based on the highest similarity. Finally, the proposed approach was evaluated by comparing the results with the human recommendations using experts. As a result, the F1 score of the proposed approach was .81. In the end, we faced some limitations in this research, including the lack of research that contributes to users modeling based on LinkedIn profiles in general, contributing to providing customized recommendations based on LinkedIn profiles in particular.

## VI. FUTURE WORK

For future work, we will consider implementing this approach with some enhancements, including:

- Implementing the proposed algorithm in Arabic.
- Modeling users utilizing other LinkedIn sections, such as education and experience, and then comparing them.
- Evaluating the proposed algorithm after including it in one of the MOOC platforms.

## REFERENCES

[1] S. Assami, N. Daoudi, R. Ajhoun, Personalization criteria for enhancing learner engagement in MOOC platforms, IEEE Glob. Eng. Educ. Conf. EDUCON. 2018-April (2018) 1265–1272. https://doi.org/10.1109/EDUCON.2018.8363375.

[2] A.A. Kardan, A. Narimani, F. Ataiefard, A Hybrid Approach for Thread Recommendation in MOOC Forums, Waset.Org. 11 (2017) 2175–2181. https://www.waset.org/publications/10007978.

[3] K. Julia, V.R. Peter, K. Marco, Educational scalability in MOOCs: Analysing instructional designs to find best practices, Comput. Educ. 161 (2021) 104054. https://doi.org/10.1016/j.compedu.2020.104054.

[4] N. DEMIRCI, What is Massive Open Online Courses (MOOCs) and What is promising us for learning?: A Review-evaluative Article about MOOCs, Necatibey Eğitim Fakültesi Elektron. Fen ve Mat. Eğitimi Derg. 8 (2013) 231–256. https://doi.org/10.12973/nefmed.2014.8.1.a10.

[5] S. Blum-Smith, M.M. Yurkofsky, K. Brennan, Stepping back and stepping in: Facilitating learner-centered experiences in MOOCs, Comput. Educ. 160 (2021) 104042. https://doi.org/10.1016/j.compedu.2020.104042.

[6] N. Jitpaisarnwattana, H. Reinders, P. Darasawang, Language MOOCs: An Expanding Field, Technol. Lang. Teach. Learn. 1 (2019) 21–32. https://doi.org/10.29140/tltl.v1n1.142.

[7] D. Rădoiu, Organization and Constraints of a Recommender System for Moocs, Sci. Bull. Univ. Tîrgu Mureş. 11 (2014) 2286–3184.

[8] X. Yang, Y. Guo, Y. Liu, H. Steck, A survey of collaborative filtering based social recommender systems, Comput. Commun. 41 (2014) 1–10. https://doi.org/10.1016/j.comcom.2013.06.009.

[9] V. Pourheidari, E.S. Mollashahi, J. Vassileva, R. Deters, Recommender System based on Extracted Data from Different Social Media. A Study of Twitter and LinkedIn, 2018 IEEE 9th Annu. Inf. Technol. Electron. Mob. Commun. Conf. IEMCON 2018. (2019) 215–222. https://doi.org/10.1109/IEMCON.2018.8614793.

[10] J. He, W.W. Chu, A Social Network-Based Recommender System (SNRS), Encycl. Soc. Netw. Anal. Min. (2018) 2699–2699. https://doi.org/10.1007/978-1-4939-7131-2_101173.

[11] F. Ricci, L. Rokach, B. Shapira, P.B. Kantor, Recommender Systems Handbook, Springer US, 1989. https://doi.org/10.1017/CBO9781107415324.004.

[12] K. Abhinav, V. Subramanian, A. Dubey, P. Bhat, A.D. Venkat, LeCoRe : A Framework for Modeling Learner ' s preference, Educ. Data Min. Conf. (2018).

[13] X. Li, T. Wang, H. Wang, J. Tang, Understanding User Interests Acquisition in Personalized Online Course Recommendation, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 11268 LNCS (2018) 230–242. https://doi.org/10.1007/978-3-030-01298-4_20.

[14] K. Alzahrani, M. Maccawy, A Hybrid Personalization Model for Searching Multiple MOOCs, (2019) 42–52.

[15] A.K. Id, K. Lundqvist, A.Y. Id, M.A. Ghzanfar, Novel online Recommendation algorithm for Massive Open Online Courses ( NoR-MOOCs ), (2021) 1–21.

[16] G. Piao, J.G. Breslin, Analyzing MOOC entries of professionals on linked in for user modeling and personalized MOOC recommendations, UMAP 2016 - Proc. 2016 Conf. User Model. Adapt. Pers. (2016) 291–292. https://doi.org/10.1145/2930238.2930264.

[17] X. Jing, J. Tang, Guess you like: Course recommendation in MOOCs, Proc. - 2017 IEEE/WIC/ACM Int. Conf. Web Intell. WI 2017. (2017) 783–789. https://doi.org/10.1145/3106426.3106478.

[18] L. Chen, D. Ifenthaler, The Adoption of Intelligent and Virtual Teams in Online Entrepreneurship Education Courses, (n.d.) 8–11.

[19] R. Huang, R. Lu, Research on Content-based MOOC Recommender Model, 2018 5th Int. Conf. Syst. Informatics, ICSAI 2018. (2019) 676–681. https://doi.org/10.1109/ICSAI.2018.8599503.

[20] H. Zhang, T. Huang, Z. Lv, S. Liu, H. Yang, MOOCRC: A Highly Accurate Resource Recommendation Model for Use in MOOC Environments, Mob. Networks Appl. 24 (2019) 34–46. https://doi.org/10.1007/s11036-018-1131-y.

[21] A. Chanaa, N.E. El Faddouli, Context-aware factorization machine for recommendation in Massive Open Online Courses(MOOCs), 2019 Int. Conf. Wirel. Technol. Embed. Intell. Syst. WITS 2019. (2019) 1–6. https://doi.org/10.1109/WITS.2019.8723670.

[22] M. Bastian, M. Hayes, W. Vaughan, S. Shah, P. Skomoroch, H. Kim, Linked in skills: Large-scale topic extraction and inference, RecSys 2014 - Proc. 8th ACM Conf. Recomm. Syst. (2014) 1–8. https://doi.org/10.1145/2645710.2645729.

[23] N. van de Ven, A. Bogaert, A. Serlie, M.J. Brandt, J.J.A. Denissen, Personality perception based on LinkedIn profiles, J. Manag. Psychol. 32 (2017) 418–429. https://doi.org/10.1108/JMP-07-2016-0220.

[24] K. Dai, C.G. Nespereira, A.F. Vilas, R.P.D. Redondo, Scraping and Clustering Techniques for the Characterization of Linkedin Profiles, (2015). https://doi.org/10.5121/csit.2015.50101.

[25] K. Dai, A.F. Vilas, R.P.D. Redondo, A New MOOCs' Recommendation Framework based on LinkedIn Data, (2017) 13–19. https://doi.org/10.1007/978-981-10-2419-1.

[26] L.D. Kumalasari, A. Susanto, Recommendation System of Information Technology Jobs using Collaborative Filtering Method Based on LinkedIn Skills Endorsement, 5 (2020) 35–39. https://doi.org/10.24167/sisforma.

[27] O. Korableva, T. Durand, O. Kalimullina, I. Stepanova, Studying user satisfaction with the MOOC platform interfaces using the example of coursera and open education platforms, ACM Int. Conf. Proceeding Ser. (2019) 26–30. https://doi.org/10.1145/3322134.3322139.

[28] S.L. Schneider, The classification of education in surveys: a generalized framework for ex-post harmonization, Springer Netherlands, 2021. https://doi.org/10.1007/s11135-021-01101-1.

[29] H. Wilde, V. Knight, J. Gillard, A novel initialisation based on hospital-resident assignment for the k-modes algorithm, (2020) 1–24. http://arxiv.org/abs/2002.02701.

[30] S. Qaiser, R. Ali, Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents, Int. J. Comput. Appl. 181 (2018) 25–29. https://doi.org/10.5120/ijca2018917395.

[31] D. Wang, Y. Liang, D. Xu, X. Feng, R. Guan, A content-based recommender system for computer science publications, Knowledge-Based Syst. 157 (2018) 1–9. https://doi.org/10.1016/j.knosys.2018.05.001.

[32] H.J. Kim, T.S. Kim, S.Y. Sohn, Recommendation of startups as technology cooperation candidates from the perspectives of similarity and potential: A deep learning approach, Decis. Support Syst. 130 (2020) 113229. https://doi.org/10.1016/j.dss.2019.113229.