Received: November 1, 2017
Revision received: May 2, 2018
Accepted: May 5, 2018

www.estp.com.tr
DOI 10.12738/estp.2018.5.088 ♦ October 2018 ♦ 18(5) ♦ 1876-1886

Research Article

Design and Implementation of Employment Recommendation Service Platform for College Graduates

Liwei Gu[1]
Hebei Normal University of Science & Technology

Zhe Zhou[2]
Hebei Normal University of Science & Technology

Yulian Zhu[3]
Hebei Normal University of Science & Technology

Abstract

In recent years, the university enrolment expansion policy has led to the continuously increasing college graduates. Every year, a large number of graduates flock for massive business recruitment information, while it costs much time and effort to find the suitable jobs in such huge recruitment information. In this context, this paper proposed an employment recommendation platform for college graduates under the background of big data. The platform calculates the similarity index between graduates and recruiting enterprises by the SimRank algorithm and K-Means algorithm successively. Then, the application index was obtained according to the PageRank algorithm. Finally, the similarity index and application index were matched to get the recommended ranking weight of enterprises. In this way, the enterprises ranking among the first can be recommended to corresponding graduates. This platform has been tested and can achieve the initial purpose, which can be used to provide graduates with scientific and rational recommendation services. It is of high practical value as it can improve application successful rate and reduce the time cost for looking a job.

Keywords

K-Means Algorithm • Recommendation Services • Recommendation Ranking • Simrank Algorithm

[1] **Correspondence to:** Liwei Gu, College of Education, Hebei Normal University of Science & Technology, Qinhuangdao 066004, China. Email: gu.liwei@163.com
[2] College of Education, Hebei Normal University of Science & Technology, Qinhuangdao 066004, China. Email: 30451344@qq.com
[3] College of Education, Hebei Normal University of Science & Technology, Qinhuangdao 066004, China. Email: 121127127@qq.com

Citation: Gu, L. W. & Zhou, Z. (2018). Design and Implementation of Employment Recommendation Service Platform for College Graduates. *Educational Sciences: Theory & Practice, 18(5),* 1876-1886. http://dx.doi.org/10.12738/estp.2018.5.088

In recent years, due to the general enrolment expansion of colleges and universities, the number of college graduates has continued to grow, while the downturn in the market economy poses a severe employment situation to each graduate. In the face of a large number of graduates and massive corporate recruitment information, it is essential to provide timely and effective employment guidance and referral services for each graduate. However, the current employment service platforms of the universities only publish the enterprise recruitment information, but cannot offer effective recommendation to graduates. And it is time and energy consuming for graduates to target the favourable jobs in a large number of recruitment information with low successful rate (Liu, Ke, Lee & Lee, 2008; Kuo, Liao & Tu, 2005; Zahra *et al*., 2015). In this context, this paper proposed an employment recommendation platform for college graduates under the background of big data, which can make recommendations scientifically and properly, improve the successfully rate, and reduce time costs.

Overview of Relevant Techniques

SimRank Algorithm

Assuming that there is a similar relationship between any two objects in a set, and their associated objects are also similar, then the SimRank algorithm can be used to calculate the similarity between any two objects (Bilge & Polat, 2013; Kathuria, Jansen, Hafernik & Spink, 2010; Kurucz & Szelenyi, 2010). If you need to calculate a large number of samples, the iterative method is needed in the operation process to complete the similarity calculation. First of all, the similarity between any two objects is initialized according to Formula (1):

$$R_0(a,b) = \begin{cases} 0 a \neq b \\ 1 a = b \end{cases} \tag{1}$$

In the formula, (a, b) represents the similarity between a and b at the 0th iteration. And then the next iteration is done on the basis of the result of each last iteration. The corresponding formula data will also be updated as Formula (2):

$$R_{k+1}(a,b) = v \begin{cases} 1, \text{if a} = b \\ \frac{c}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} R_k\left(I_k(a), I_j(b)\right), \text{if a} \neq b \end{cases} \tag{2}$$

Keep iterations until the final Rk + 1 (a, b) is obtained, and the final data is the similarity index of the two objects of a and b.

K-Means algorithm

The K-Means algorithm as a clustering method, uses a division method also known as the class center, which is based on the record with a representative feature in a given data set. Then the class center is used to divide the records of the data set (Harris *et al*., 2014; Baert *et al*., 2017; Kim & Ahn, 2008; Tzortzis, Likas & Tzortzis, 2014). Suppose that a given data set has N records, where the K records have representative characteristics and are selected as the initial clustering center, the distance between the remaining records in the data set and the records of these clustering centers shall be calculated. And according to the nearest distance

principle, these records can respectively be divided into the corresponding clusters. And then based on the clusters, the average value of each cluster is calculated to update the clustering center. Keep repeating the above steps until each cluster converges (Sobiech & Dierking, 2013; Mosél & Goheen, 2010; Goheen & Mosél, 2010).

PageRank Algorithm

The PageRank algorithm, as part of the Google ranking algorithm, assigns each page a value that measures its importance and applies it to the sorting of the search results. The formula is as follows:

$$PR(A) = (1-d) + d \bullet \sum_{i=1}^{n} \frac{PR(T_i)}{C(T_i)} \quad (3)$$

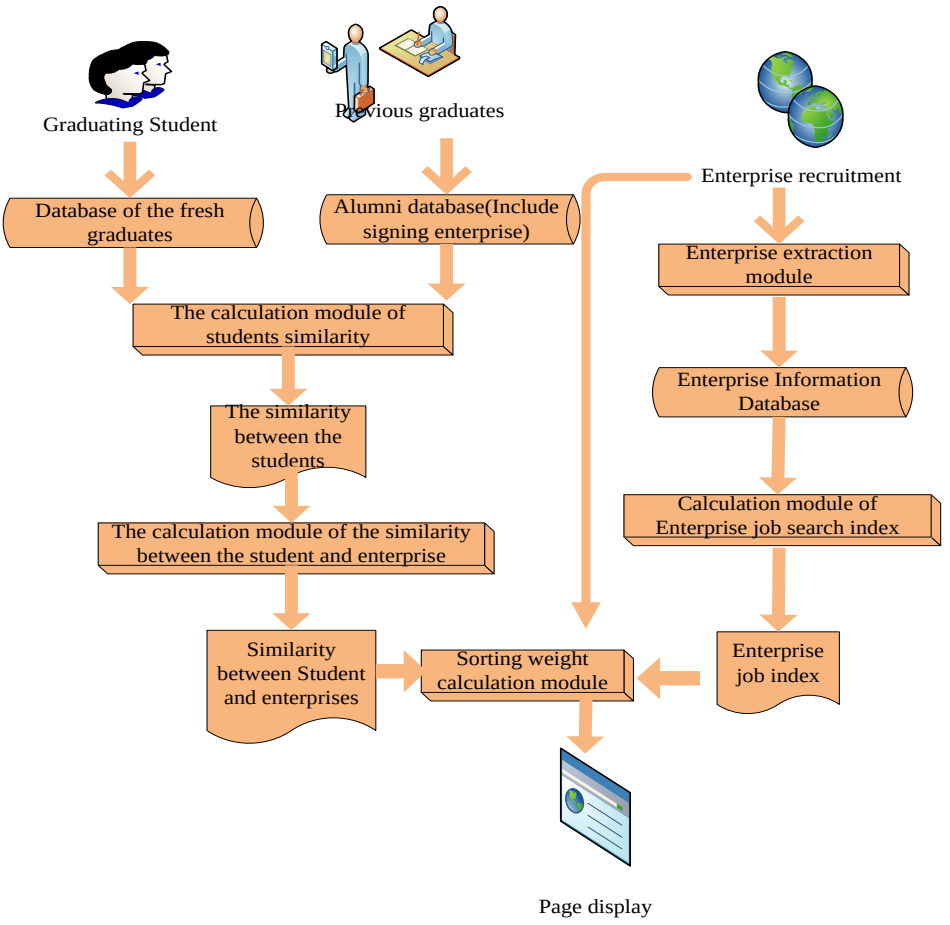

Figure 1. Framework of Recommendation System for college graduates.

PR (A) represents the PageRank value of Page A, PR (Ti) the PageRank value of Page Ti that links Page A, and C (Ti) the number of outbound links of Page Ti. d is the damping coefficient. Generally speaking, it goes the rule of 0<d<l (Mosel & Goheen, 1958; Knouse, 2011; Shibata, Oka, Nakamura & Muraoka, 2009).

Figure 2. Hardware Configurations of server for developing system.

Overview of the Platform

The overall design framework of the platform

The design framework of the platform is shown in Figure 1:

It can be seen from the figure that the design module of the system mainly includes the data preprocessing module, the enterprise information extraction module, the student similarity calculation module, the student cluster analysis module, the student-enterprise similarity calculation module, the enterprise contracting index calculation module, and the sorting weight calculation module.

Introduction to modules

The data pre-processing module is mainly for the pre-treatment of college student data. For example, the student titles are unified as "non-cadre" and "class cadre" and "department cadre", which standardize the data and ensures the only meaning of each data.

The enterprise information extraction module is responsible for the extraction of key content and the standardization of data concerning all the enterprise recruitment information, such as the basic information and recruitment requirements of enterprises. Then these key contents are stored into the "enterprise database".

The rest calculation modules are the key technology modules in the design process of the platform. In these modules, the SimRank algorithm and the K-Means algorithm are successively used to calculate the similarity index between the graduates and the recruiting

irms. Then, according to the PageRank algorithm, the application index of the firm is obtained. At last, the similarity index and the application index are matched to obtain the final ranking weight.

System environment

The whole process of the development of the platform requires large-scale database storage and a lot of data operations. In order to ensure the speed of the platform, the server hardware is configured as shown in Figure 2.

Technical Implementation of Key Modules

Students similarity calculation based on Simrank algorithm

The prerequisite of calculating the similarity is to find out the similar relationship between students. According to the core idea of SimRank algorithm, the characteristics and properties of students is turned into an interrelated binary relationship as shown in Table 1. And then this binary relationship is converted into the form of an undirected graph, as shown in Figure 3.

Table 1
Binary Relation between Students and their Feature Attribute

Student	Feature	Student	Feature
Wang Fei	Computer	Li Ming	Electronic message
Wang Fei	The school rewards	Li Ming	No prize
Wang Fei	Liao Ning	Li Ming	Shan Dong
Wang Fei	Party member	Li Ming	League member
Wang Fei	English Cet 6	Li Ming	English Cet 4
Wang Fei	Man	Li Ming	Man
Wang Fei	Undergraduate	Li Ming	Undergraduate
Wang Fei	School cadre	Li Ming	Not a cadre
WangFei	2007	Li Ming	2007

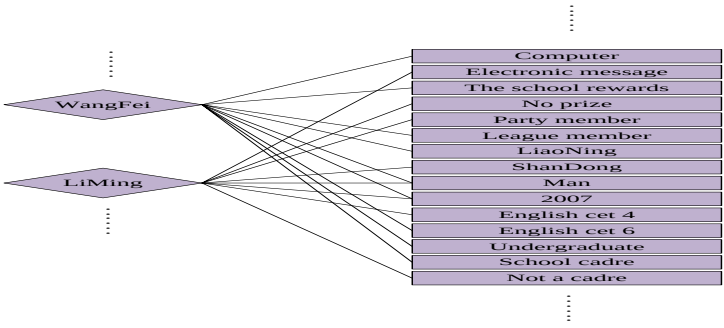

Figure 3. Related Undirected Graph of table 1.

In Figure 3, based on the SimRank algorithm, the similarity between students is calculated as Formula (4), and the similarity between the characteristics is obtained via Formula (5):

$$S(D_1, D_2) = \begin{cases} 1, \text{if } D_1 = D_2 \\ \frac{C}{|I(D_1)||I(D_2)|} \sum_{i=1}^{I(D_1)} \sum_{j=1}^{|I(D_2)|} S(I_i(D_1), I_j(D_2)), \text{if } D_1 \neq D_2 \end{cases} \quad (4)$$

$$S(E_1, E_2) = \begin{cases} 1, \text{if } E_1 = E_2 \\ \frac{C}{|I(E_1)||I(E_2)|} \sum_{i=1}^{I(E_1)} \sum_{j=1}^{|I(E_2)|} S(I_i(E_1), I_j(E_2)), \text{if } E_1 \neq E_2 \end{cases} \quad (5)$$

By the given formula and the correlation graph, taking two students "Wang Fei" and "Li Ming" in Figure 3 as an example, the following formula of similarity can be drawn:

S (Wang Fei, Li Ming) = [S (computer, electronics) + S (school award, no awards) + S (Party members, League members) +. ... + S (school cadre, non-cadre)] ×C/(9×9)

As for the two characteristic attributes "computer" and "electronics" in Figure 3, their similarity can be calculated as:

S (computer, electronics) = [... + S (Wang Fei, Li Ming) + ...] × C / (the number of students connected to "computer" × the number of students connected to "electronics")

The calculation process of the above formulas finally leads to the process of solving the equation set. In order to avoid the huge computation amount in solving the equation set, the "iteration" method is used to assist the calculation, with the process as follows:

(1) Initialize the data at the beginning of the operation. For example, the similarity between any two students is initialized to 0, and the similarity between each student and himself is initialized to 1. The characteristic attribute can also be initialized in this way.

(2) Alternately use the formula (4) and (5) for calculation, during which the similarity data will be updated continuously;

(3) Repeat the formula alternately until the convergence.

K–means based analysis of student clusters

The similarity has been calculated via the SimRank algorithm. The next step is to analyze the student clusters based on the K-Means algorithm, with details as follows:

(1) Select K students and identify them as the clustering center;

(2) Compare the rest of the students with the K students in the cluster center regarding the similarity, and use the minimization principle to "gather" the other students in the clustering center determined in (1);

(3) After the clustering results are obtained, reselect the clustering center for the next iteration;

(4) Iteratively executes Step (2) and (3) N times;

(5) Complete clustering, and label the data and clustering center.

The similarity between graduates and enterprise

Set fresh graduates as i, former graduates as j, the contracting enterprise where former graduates j work as w, and the similarity between any student in the fresh graduate database and all the former graduates the most similar to the fresh graduate in the former graduate database as Sim_{exa}, leading to Simexa (i, j) = Simstu (i, j).

Since each of the former graduates has the only contracted enterprise, it can simply put that the similarity between fresh graduates i and the contracted enterprises of former graduates j equals to the the similarity between fresh graduates i and former graduates j, which is Simexa (i, j). If Sim'co is used to represent the similarity between fresh graduates and enterprises w, then Sim'co (i, w) = Simexa (i, j) is obtained.

In applications, it is necessary to revise the definition of the similarity between former graduates and enterprises:

$$Sim_{co}(i,w) = \overline{Sim'_{co}(i,w)} + (|Sim'_{co}(i,w)| - 1) \bullet \alpha \quad (6)$$

$\overline{Sim'_{co}(i,w)}$ stands for the average value of all similarity data between fresh graduates i and enterprises w. $|Sim'_{co}(i,w)|$ represents the number of similarities calculated between fresh graduates i and enterprises w, also the number of former graduates that are corresponding to fresh graduates and contracted with enterprises w. α refers to the factors that influence the cumulative number of contracting students of enterprises, and can be set according to the actual situation of different colleges and universities.

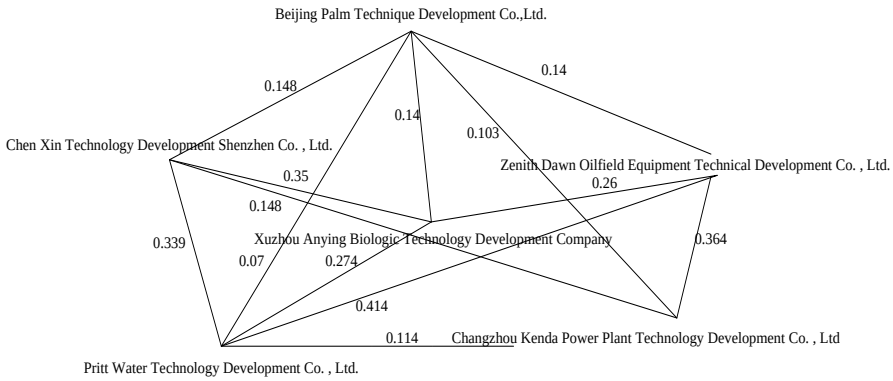

Figure 4. Sub-graph of Graph G.

Enterprise application index calculation based on pagerank algorithm

According to the characteristics of the platform, this paper uses the random surfer algorithm of PageRank to set each enterprise as a node V, the similarity based on the basic characteristics of the enterprises as the edge. And each edge has a weight E. Then the weighted undirected graph $G_{<V, E>}$ is built, with the sub graph of Figure G shown in Figure 4.

Table 2
Statements of Basic Feature of Enterprises

Company name: *****
Take up an occupation: spaceflight communication software
Location: Shenzhen Beijing Shanghai Xian
Recruitment major: computers Software physics English Electronics
Education background of recruitment: Undergraduate Master Doctor
Enterprise property: State-owned enterprise

According to the basic characteristics in Table 2, Formula (7) is used to calculate the similarity between any two firms in the "enterprise database", that is the weight $E_{(i,j)}$ of the connected edge of any two nodes V_i and V_j on Graph G.

$$E_{(i,j)} = \sum_{k=1}^{n} \left(\frac{\left|T_i^k \cap T_j^k\right|}{\left|T_i^k \cup T_j^k\right|} * \sigma_k\right) \tag{7}$$

In Formula (7), when k values from 1 to 5, T^k respectively represents the value set of the five categories of characteristics of "industry", "area", "major", "educational background" and "corporate nature". σ_k represents the weight of the kth characteristic type in the calculation of the similarity between firms. For different colleges and universities, the value selected of σ_k can be different values. This

paper set σ_k as 0.27, σ_B as 0.16, σ_C as 0.31, σ_D as 0.14, and σ_E as 0.12.

In Figure G, according to the relations of the page links, start from any node, and randomly surfer to other nodes. When it is at the converging point, the nodes with higher points are more important. The PageRank algorithm can be used to finally calculate the PR value for each node, which reflects not only the degree of concern for the node, but also the possibility of finding the node through other nodes via the link relations. Then the value of enterprise "application index" (PR) is obtained via the PageRank algorithm from the formula as follows:

$$PR(A) = (1-d) + d \bullet \sum_{i=1}^{n} \frac{PR(T_i)}{C(T_i)} \tag{8}$$

Based on the weighted undirected graph G mentioned in Figure 4, Formula (8) can be interpreted as follows:

(1) PR (A) represents the "application index" of Firm A (A node in Figure G);

(2) PR (T_i) represents the "application index" of the T_i node (Enterprise T_i) connected to the A node in graph G;

(3) C (T_i) represents the out-degree of the Ti node in graph G. Since Graph G is undirected, so the out-degree and in-degree of node Ti are even and equal to the degree of the node;

(4) d is the damping coefficient with 0 <d <1. According to the PageRank algorithm, the value of d is set as 0.85.

The iterative method is used to calculate the PR value. Considering the running time and actual needs of the module, the convergence threshold that controls the iteration process is set as 0.001. That is, if the changes of the iteration result is less than 0.001 compared with the previous iteration result, the iteration is terminated.

Calculation of the ranking weight

Based on the Sim_{co} of the similarity the students and firms, and the "application index" PR, the weights for ranking can be calculated by Formula (9).

$$W(i,w) = Sim_{CO}(i,w) * \lambda + \frac{PR(w)}{PR_{max}} \quad (9)$$

W (i, w) represents the sorting weight of the enterprise w in the recommended enterprise of the student i. In line with the distribution of the “application index” PR calculated, PR_{max} is defined as the maximum value of the PR values of all enterprises.

Platform Test and Operations Analysis

After the completion of the design, the platform is tested for the application. The test shows that when a graduate login in the platform, the platform will display the 20 recommended enterprises ranking in line with the W values in the recommendation column of the personal interface. The students can click the name of the enterprise to obtain the company's recruitment information. We can further understand the actual effect of the platform through the following examples.

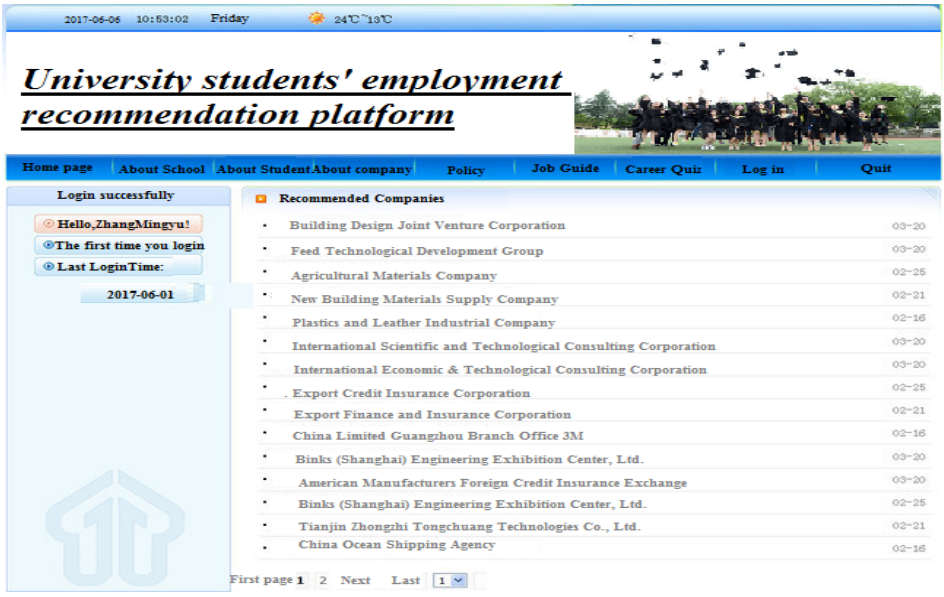

Figure 5. Display page 1 of recommendation results.

For the graduate "Zhang Mingyu" listed in Fig. 5, his real information is: "computer major", "2009", "master", "Communist Party member", "male", "CET-6", "Liaoning Dalian "," Non-cadre "," no prizes",

The graduate actually signed contract with the "architectural design joint venture", which ranks the first in the 20 recommended enterprises provided to him. Moreover, after the analysis by the employment guidance department, all of the top 20 companies recommended by the system can be regarded as the potential contracting enterprises for the graduate.

Some students were invited to try the platform. From their feedback, it could be concluded that the enterprise list recommended by the platform basically meets the expectations of the students with a relatively high accuracy. And with this platform, graduates can save much more time and energy in finding a job than searching in the large amount of recruitment information, which can improve the successful rate of looking for a job and reduce the application costs to a certain extent. At the same time, it also provides references for students who have no clear career objectives.

Conclusion

(1) The platform was designed based on the calculation of the student similarity via the SimRank algorithm, and that of the enterprise application index via the PageRank algorithm;

(2) The K-Means was used to analyze the clustering of former graduates, leading more rational results of the similarity between students and enterprises.

(3) Proved by the test, the platform can provide scientific and reasonable job recommended services for graduates to improve application successful rate and reduce the time costs, with high practical value.

References

Baert, Q., Caron, A.C., Morge, M., Routier, J.C. (2017). Fair task allocation for large data sets analysis, *Revue d'Intelligence Artificielle, 31(4),* 401-426. http://dx.doi. org/10.3166/RIA.31.401-426

Bilge A., & Polat, H. (2013). A scalable privacy-preserving recommendation scheme via bisecting k -means clustering. *Information Processing & Management*, *49*(4), 912-927.

Goheen, H. W., & Mosél, J. N. (2010). Validity of the employment recommendation questionnaire: ii. Comparison with field investigations. *Personnel Psychology*, *12*(2), 297-301. http://dx.doi. org/10.1111/j.1744-6570.1959.tb00811.x

Harris, J. R., Hannon, P. A., Beresford, S. A. A., Linnan, L. A., & Mclellan, D. L. (2014). Health promotion in smaller workplaces in the United States. *Annual Review of Public Health*, *35*(35), 327. http://dx.doi. org/10.1146/annurev-publhealth-032013-182416

Kathuria, A., Jansen, B. J., Hafernik, C., & Spink, A. (2010). Classifying the user intent of web queries using k-means clustering. *Internet Research*, *20*(5), 563-581. http://dx.doi. org/10.1108/10662241011084112

Kim, K. J., & Ahn, H. (2008). A recommender system using ga k-means clustering in an online shopping market. *Expert Systems with Applications, 34*(2), 1200-1209. http://dx.doi. org/10.1016/j.eswa.2006.12.025

Knouse, S. B. (2011). Improving the letter of recommendation. *Journal of Employment Counseling, 31*(3), 105-109.

Kuo, R. J., Liao, J. L., & Tu, C. (2005). Integration of art2 neural network and genetic k-means algorithm for analyzing web browsing paths in electronic commerce. *Decision Support Systems, 40*(2), 355-374.

Kurucz, I., & Szelenyi, I. (2006). Current animal models of bronchial asthma. *Current Pharmaceutical Design, 12*(25), 3175. http://dx.doi. org/10.2174/138161206778194169

Liu, D. R., Ke, C. K., Lee, J.Y., & Lee, C. F. (2008). Knowledge maps for composite e-services: a mining-based system platform coupling with recommendations. *Expert Systems with Applications, 34*(1), 700-716. http://dx.doi. org/10.1016/j.eswa.2006.10.005

Mosel, J. N., & Goheen, H. W. (1958). The validity of the employment recommendation questionnaire in personnel selection: I skilled traders. *Personnel Psychology, 11*(4), 481-490. http://dx.doi. org/10.1111/j.1744-6570.1958.tb00034.x

Mosél, J. N., & Goheen, H. W. (2010). The employment recommendation questionnaire: iii. Validity of different types of references. *Personnel Psychology, 12*(3), 469-478. http://dx.doi. org/10.1111/j.1744-6570.1959.tb01338.x

Shibata, A., Oka, K., Nakamura, Y., & Muraoka, I. (2009). Prevalence and demographic correlates of meeting the physical activity recommendation among Japanese adults. *Journal of Physical Activity & Health, 6*(1), 24-32. http://dx.doi. org/10.1037/a0013822

Sobiech, J., & Dierking, W. (2013). Observing lake- and river-ice decay with sar: advantages and limitations of the unsupervised k-means classification approach. *Annals of Glaciology, 54*(62), 65-72.

Tzortzis, G., Likas, A., & Tzortzis, G. (2014). The minmax k-means clustering algorithm. *Pattern Recognition, 47*(7), 2505-2516. http://dx.doi. org/10.1016/j.patcog.2014.01.015

Zahra, S., Ghazanfar, M. A., Khalid, A., Azam, M. A., Prugel-Bennett, A., & Prugel-Bennett, A. (2015). Novel centroid selection approaches for Kmeans-clustering based recommender systems. *Information Sciences, 320*(C), 156-189. http://dx.doi. org/10.1016/j.ins.2015.03.062