Link Analysis in Employment Data Set to Improve Learning Outcomes for IT Programmes

Kadhim B. Swadi Al-Janabi
Mathematics Department, College of Mathematics and Computer Sciences
University of Kufa
kadhimbs@yahoo.com

ABSTRACT
This paper presents an approach for analyzing data of the Information Technology graduates according to the employability knowledge areas in order to predict feedback recommendations to improve the IT programmes teaching and learning resources and processes towards the improvement of the programme learning outcomes. The approach is based on features (knowledge areas) extracted from logged data for employment and university graduates. Link analysis is an efficient approach to study the correlation and relationships between different attributes that highly affect jobs in IT market, including different skills areas in both the market and the programme curriculum, and it gives good weighted evaluation for these knowledge areas. The link analysis shows great relationship and associations between these attributes (Student Performance in Bachelor degree, analytical and development skills, Programming skills (Java, C++, C#, etc), practical skills, communication skills, and training and certificates) and the market demands. Data set from IT market and university records is used to create and test the model. WEKA was used as a software for mining tasks.

Categories and Subject Descriptors
Database Applications, Data Mining
Keywords
Data Mining, Classification, Association, Link Analysis

1. INTRODUCTION

Curriculum contents, design, and organization for IT programmes are based mainly on the following resources:
IEEE/ACM recommendations [1].
Quality Assurance Recommendations [2].
Local, Regional, and Global market demands.
Local organizations regulations.
Market demands and employability knowledge areas represent common criterion for the resources mentioned above. In this paper we are interesting in the data received from the IT market and employment feedback that gives us the knowledge area as main criteria for getting a job within six months from the graduation date in their specialization. The proposed framework in this paper tries to answer the following questions:

1. Can we find out the attribute(s) that represent crucial factors in getting jobs in IT market? If yes, can we find out the weight for each attribute?
2. Can we find the relationships between each of these attributes and the jobs in IT market? If yes, can we concentrate on the most effective relationships?
3. Can we apply link analysis as a data mining technique to find such relationships and the weight for each attribute?
4. How can we reflect the knowledge areas needed in IT market in IT departments curriculums?

Data Mining (DM) tasks can be classified into the following main topics [3,5].

1. Classes: Stored data are used to locate objects in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials. In the proposed framework this task can be used in classifying the IT graduates into groups according to their knowledge and skills.
2. Clusters: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities. The most effective knowledge areas in IT jobs fit in this category.
3. Associations and Link Analysis: Data can be mined to identify associations. The Smoker-Cancer example is an example of associative mining. What are the relationships between different knowledge areas and job classes?
4. Sequential patterns: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.
5. Prediction: Data can be used to find out their trends and behavior and to predict their future behavior according to the historical data stored in data warehouse, for example, what will be the gold prices for the next two months based on the historical data. Out Can we find out the trend and predict the IT market behavior for the next few months?

Our aim is to find relations between IT graduate skills (program outcomes) and employment knowledge areas, this will be used as feedback to improve teaching and learning process.

Many previous works[3,4,5,6,7,8,9,10,13,14] tried to improve the student performance and skills through the enhancement of different models inside the teaching and learning body, anyhow link analysis presented in this paper tried to enhance employability knowledge required in programme curriculum and teaching and learning methodologies.

2. DATA COLLECTING AND PREPROCESSING

As data set, we selected IT graduates from different IT departments and Programmes. The number of graduates selected was 105. The extracted features (Attributes) that represent the main factors for employability are given below:

1. Student Performance in Bachelor degree (Overall Average)
2. Analytical and development skills.
3. Programming skills (Java, C++, C#, etc)
4. Practical skills(Using software and tools)
5. Communication skills (English language, Presentations, Demos, etc)
6. Training and certificates (Oracle, Java, Cisco, etc)

The survey conducted in the IT market showed that the knowledge areas mentioned above have different weights in decision making, accordingly, Table(I) gives an estimate for them. Binning technique[3] is used to get a categorical estimate for each knowledge area in order to produce clear groups of weights. Accordingly, the graduate scores distribution is given in Table (II).

Table(I). Weight distribution corresponding to employability knowledge areas

Knowledge Area	Category	Weight
Performance	High Medium Low	25 15 10
Development	Yes No	25 10
Programming	Excellent Satisfactory Poor	15 10 5
Practical	Yes No	15 5
Communication	Excellent Medium Bad	10 5 0
Training	Yes No	10 0

Table (II). Number of graduates according to the employability score distribution

Score	30 – 40	40 – 50	50 – 60	60 – 70	70 – 80	80 – 90	90 - 100
Number of Graduates	1	13	29	33	23	5	1

Figures 1 shows line chart representation of Table(I), whereas figure 2 represents score categories versus job classes..

Figure 1. Graph of Distribution of Employability Scores.

Figure 2. Graph of Score Category vs. Job Distribution

The collected data were preprocessed for the following reasons:

- Real world data are usually Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data.
- Noisy data: containing errors or outliers
- Data are Inconsistent: containing discrepancies in codes or names

So that data needed to be preprocessed using the following tasks:

- Data cleaning: fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.
- Data integration: using multiple databases, data cubes, or files. Since our data were collected from university records and the employability markets, the data need to be integrated in one unit.
- Data transformation: normalization and aggregation. Scores and ranges for the graduates were normalized and aggregated to get better knowledge.

Data reduction: reducing the volume but producing the same or similar analytical results. Samples of data that represent different categories of graduates were used to get efficient analytical algorithms.

- Data discretization: part of data reduction, replacing numerical attributes with nominal ones. Graduate scores were replaced with nominal values to be suitable for processing.

3. Proposed Framework

The proposed framework and architecture is shown in figure 3. The components are:

1. Databases from which the data are collected and this includes university graduation records and the employability database from the IT market.
2. Preprocessing stage in which the collected data are Extracted, Transformed, and Loaded (ETL) into the required repository suitable for mining and statistical purposes.
3. Applying mining task represented by Link Analysis to find out relationships of different attributes (Knowledge and skills) with the job classes in addition to the weight of each attribute.
4. Representing the results into Link Analysis Graph connecting the different attributes and weights to job classes.
5. Extracting the required knowledge from the final graph.

6-Implementing the knowledge into recommendations to improve curriculum content, design, and organization.

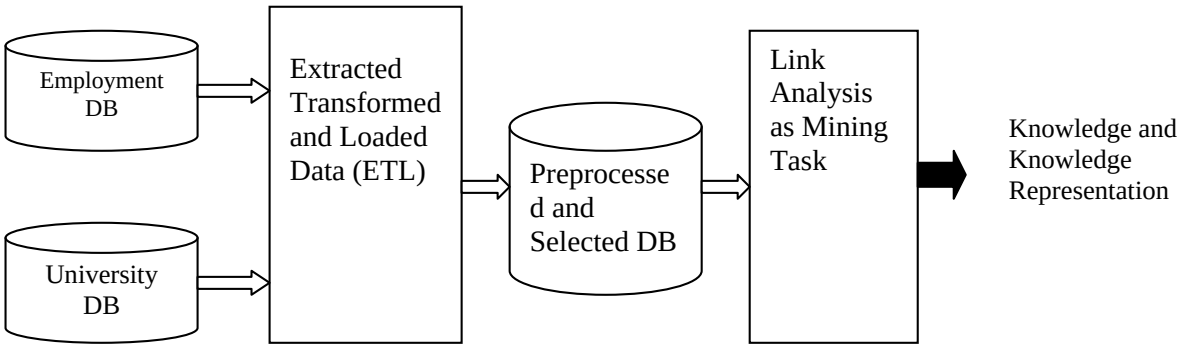


Figure 3. Proposed Framework Architecture for Link Analysis.

4. ASSOCIATIONS

Associations represents the crucial factor in mining the graduate and employability data set because it gives a good idea about what features (knowledge areas) have great effect on getting the job and what are the real relations between these knowledge areas. Apriori algorithm [3,11] is highly effective in finding out such relations. Support and Confidence are the factors that are taken into consideration when applying such algorithm, they are mentioned in equations 1 and 2 respectively[3, 5, 6].

$$Support(A \Rightarrow B) = P(A \cup B) \quad (1)$$

$$Confidence(A \Rightarrow B) = P(B \mid A) \quad (2)$$

Where $Support(A \Rightarrow B)$ refers to the probability of occurrence of attribute contents A and B to the whole data set, and $Confidence(A \Rightarrow B)$ refers to the probability of occurrence of both A and B to A data set. High values for both Support and Confidence give the indication that there exists high association between these attributes.

From the classification technique and the Decision Tree algorithm, the most effective attribute(s) is/are given using the Entropy and the Gain equations 3 and 4. The results showed that the attribute "Communication" has the maximum gain and hence it comes at the top of the decision tree. Each attribute has its own entropy and information gain and decision tree algorithms use the gain value to start splitting the tree with attribute having high gain and so on [4].

$$Entropy(S) = \sum_{i=1}^{n} - P_i \log_2 P_i$$

(3)

and the Information Gain is given in equation (2)

$$Gain(S, A) = Entropy(S) - \sum \frac{|S_v|}{|S|} Entropy(S_v) \quad (4)$$

Anyhow we are interesting in finding out the relationships between the attributes and the job class(Yes, No), and this can be achieved using Link Analysis technique.

5. Link Analysis

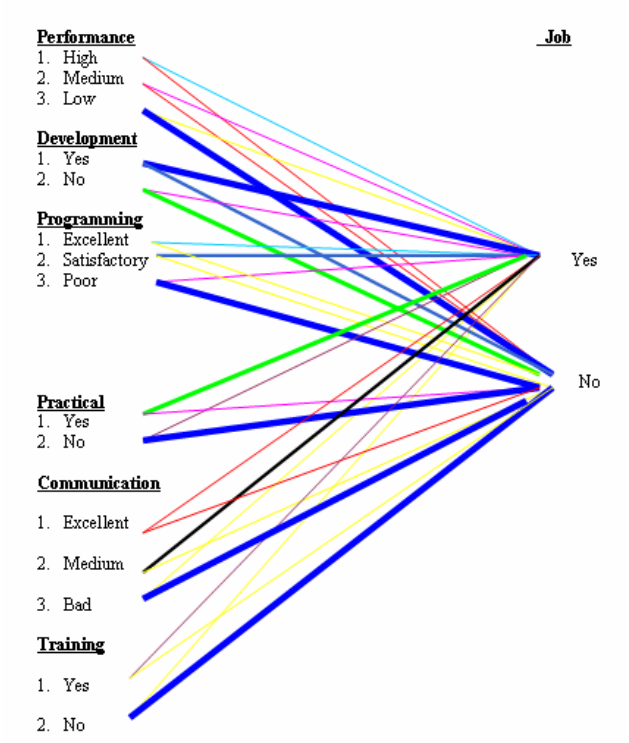
Associations is an efficient method to find out relationships between different knowledge areas and the required targeted class (job with yes or no), anyhow a much concrete concept is required to distinguish the weigh of each of the knowledge areas on the target, Link Analysis is the solution. Table(III) shows the distribution of job content(Yes, No) according to the knowledge areas previously mentioned. Table (III) below shows the distribution of job occurrence with different knowledge areas categories.

Table(III): Job Distribution according to the Knowledge Areas Categories.

Knowledge Area	Attributes	Job			
		Yes % P=0.45	No % P=0.55	Graduates# with job="Yes"	Graduates# with job="Yes"
Performance	High	0.3	0.13	14	8
	Medium	0.38	0.23	18	13
	Low	0.32	0.64	15	37
Development	Yes	0.67	0.47	32	27
	No	0.326	0.53	15	31
Programming	Excellent	0.2	0.13	9	8
	Satisfactory	0.46	0.23	22	13
	Poor	0.34	0.64	16	37
Practical	Yes	0.57	0.3	27	17
	No	0.43	0.7	21	40
Communication	Excellent	0.22	0.06	10	3
	Medium	0.48	0.3	23	17
	Bad	0.3	0.64	15	37
Training	Yes	0.36	0.366	17	21
	No	0.63	0.633	30	37

According to the data extracted from table(III), out of 105 records in the logged data set, 45% got a job and 55% didn't, and the results show the distribution of job category(yes, no) according to the knowledge area attributes, in which different attributes of knowledge areas affect the job in different manner, for example, 43% of the data set have Performance ="High", 30% with job="Yes" and 13% with job="No". The diagram shown in figure (4), gives the link analysis between these attributes and the job.

It is important to mention that the size of the lines in Link Analysis refers to high rank of associations and links between the different attributes. It is clear from figure(4) that the most important knowledge area required to get a job is development (yes) and then practical skills(yes). Mining techniques are mainly based on statistical analysis of data under consideration.



Figure(4). Link Analysis Between Different Knowledge Areas and the Job.

6. CONCLUSIONS

1. The training data and the results obtained in table (III) and figure (4) show that it is possible to predict the probability of getting a job within the estimated period according the score of a graduate in the attributes Performance, Development, Programming, Practical, Communication and Training.
2. It also showed the weight of each attributes in getting the job.
3. The results given in figure(4) showed that getting jobs in IT market fields is highly linked to the knowledge areas (Development, Practical and Communication skills).

Figure (4) gives an excellent indication about the linkage between classes of job (Yes, No) and the features. It shows that Job "No" is highly concentrated in low performance, poor programming, no practical experience and bad communication, whereas Job "Yes" is highly concentrated in graduates with good development and practical skills.

7. REFERENCES

[1] IEEE/ACM Computing Curricula, Information Technology, Volume, Version: October 2005.
www.acm.org/education/curric_vols/

[2] Quality Assurance Agency for Higher Education.
www.qaa.ac.uk

[3] Jiawei Han and Micheline Kamber "Data Mining: Concepts and Techniques" 2nd ed., Morgan Kaufmann, 2006.

[4] AlJanabi Kadhim, "Mining Employment Data Set to Improve Teachning and Learning Resources and Processes in IT Programmes, International Conference for Applied Sciences, Kufa University, 2008.

[5] B. Barros and M. F. Verdejo, 'Analysing student interaction processes in order to improve collaboration: the degree approach', International Journal of Artificial Intelligence in Education, 11, 221–241, (2000).
[6] Elena Gaudioso and Luis Talavera , Mining Student Data To Characterize Similar Behavior Groups In Unstructured Collaboration Spaces, ECAI04 workshop, 2004.
[7] Rahel Bekele Wolfgang Menzel A Bayesian Approach To Predict Performance Of A Student. Artificial Intelligence and Applications ~AIA 2005~, Innsbruck, Austria
[8] Kortemeyer G.,.Minaei-Bidgoli, B., Punch, W.F., "Association Analysis for an Online Education System", IEEE International Conference on Information Reuse and Integration (IRI-2004), Las Vegas, Nevada, Nov 2004.
[9] Kortemeyer, G., Minaei-Bidgoli, B., Punch, W.F., "Enhancing Online Learning Performance: An Application of Data Mining Methods", The 7th IASTED Internationa Conference on Computers and Advanced Technology in Education.(CATE 2004) Kauai, Hawaii, August 2004.
[10] Kashy, D.A., Kortemeyer G.,Minaei-Bidgoli, B., Punch, W.F., "Predicting Student Performance: An Application of Data Mining Methods with an educational Web-based System", (IEEE/ASEE) FIE 2003 Frontier In Education, Boulder, Colorado, Nov. 2003.
[11] D. Heckerman and M. Meila and, 'An experimental comparison of model based clustering methods', Machine Learning, 42(1/2), 9–29, (2001).
[12] M. Steinbach, P.-N.Tan and V. Kumar, Introduction to Data Mining, Addison-Wesley, 2006. ISBN: 0-321-32136-7
[13] Judy Kay, Kalina Yacef, Nicolas Maisonneuve and Osmar R. Zaiane, Mining Patterns of Events in Students' Teamwork Data, Proceedings of Educational Data Mining Workshop, held in conjunction with Intelligent Tutoring Systems (ITS), Taiwan, June 26, 2006.
[14] M. H. Dunham, Data Mining: Introductory and Advanced Topics, Prentice Hall, 2002.

الملخّص:
تُقدّمُ هذه الورقةِ البحثية نموذجا ومنهجية **(Model and Approach)** لتَحليل بياناتِ خريجي البرامج الدراسية لكليات تكنولوجيا المعلومات طبقاً للمجالات المعرفية المؤثرة في سوق العمل لغرض تحسين هذه البرامج الدراسية وتحسين مصادر التعليم والتعلم والاجراءات المطلوبة لتطوير مخرجاتها. المنهجية المقترحة في هذا البحث تستند على المجالات المعرفية المشتقة من بيانات سوق العمل وسجلات الخريجين اثناء فترة الدراسة الجامعية والتي يتم تحليلها ومعالجتها باستخدام تقنيات التنقيب عن البيانات كالتصنيف والتحليل الترابطي لهذه البيانات لغرض اكتشاف وتحديد المجالات المعرفية المؤثرة في سوق العمل . وقد استخدمت منهجية التحليل الترابطي كمنهجية ذات كفاءة عالية لدراسة هكذا حالات ودراسة وتحليل الارتباطات والعلاقات بين هذه المجالات المعرفية وسوق وفرص العمل. النتائج المستحصلة من النموذج المقترح تشير الى علاقات ترابطية مختلفة المستويات بين المجالات المعرفية المختلفة والتي تشتمل على (التحصيل الدراسي، المهارات التحليلية والتطويرية، مهارات البرمجة، المهارات العملية والتطبيقية، مهارات الاتصال، والتدريب) ومتطلبات سوق العمل. وقد تم استخدام بيانات من سوق العمل والسجلات الجامعية لبناء وتقييم النموذج المقترح، كما وتم استخدام نظام **WEKA(Waikato Environment for Knowledge Analysis)** لتحليل وتقييم النتائج والنظام المقترح.