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

<table>
<thead>
<tr>
<th>Knowledge Area</th>
<th>Category</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance</td>
<td>High</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>10</td>
</tr>
<tr>
<td>Development</td>
<td>Yes</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>10</td>
</tr>
<tr>
<td>Programming</td>
<td>Excellent</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Satisfactory</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Poor</td>
<td>5</td>
</tr>
<tr>
<td>Practical</td>
<td>Yes</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>5</td>
</tr>
<tr>
<td>Communication</td>
<td>Excellent</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Bad</td>
<td>0</td>
</tr>
<tr>
<td>Training</td>
<td>Yes</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>0</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Score</th>
<th>30 – 40</th>
<th>40 – 50</th>
<th>50 – 60</th>
<th>60 – 70</th>
<th>70 – 80</th>
<th>80 – 90</th>
<th>90 - 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Graduates</td>
<td>1</td>
<td>13</td>
<td>29</td>
<td>33</td>
<td>23</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>

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.
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 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.
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)

\[
Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} Entropy(S_i) \quad (4)
\]

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 weight 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.

![Figure 3. Proposed Framework Architecture for Link Analysis.](image-url)
Table(III): Job Distribution according to the Knowledge Areas Categories.

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

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
   www.acm.org/education/curric_vols/
   www.qaa.ac.uk


الملخص:

مقدمه: هذه الورقة البحثية نموذجاً ومنهجية (Model and Approach) لتحليل بيانات احترازي البرامج. دراسة القيود وتكنولوجيا المعلومات تسهيل للمجالات المعرفية المؤثرة في سوق العمل لعرض تحقيق هذه البرامج الدراسية وتحسين مساحة التعلم والتعليم والإجراءات المطلوبة لتطوير مخرجاتها. المنهجية المتبعة في هذا البحث تستند على المجالات المعرفية المشتركة من بيانات سوق العمل وسجلات الاحتراف أثناء فترة الدراسة الجامعية والتي يتم تحليلها وحالاتها باستخدام تقنيات التنبؤ عن البيانات كتصنيف والتحليل الترابطي لهذه البيانات لعرض اكتشاف وتحديد المجالات المعرفية المؤثرة في سوق العمل. وقد استخدمت منهجية التحليل الترابطي كمنهجية ذات كفاءة عالية لدراسة هذا حالات ودراسة وتحليل الأحداث والعلامات بين هذه المجالات المعرفية وسوق وفرص العمل. النتائج المستحيلة من النموذج المقترح تشير إلى علاقات ترابطية مختلفة المناسبة بمستويات المعرفية المختلطة ونوعية التحليلية، مهارات البرمجة، المهارات العملية والتطبيقات، مهارات الأدوات، والتدريب) ومتعلقات سوق العمل. وقد تم استخدام بيانات من سوق العمل والسجلات الجامعية لبناء وتقسيم النموذج المقترح، كما تم استخدام التحليل وتقسيم النتائج والنظم WEKA(Waikato Environment for Knowledge Analysis) للنموذج المقترح.