Lab 6: Matching Algorithms for CED Applications

Due date: May 26, 2022.

Assignment

This is a team assignment, and the last specific assignment in the course. The goal of this assignment is to have each team develop the process of matching user-created experience profiles/desired profiles with the O*NET Job Descriptions.

The overall matching process is done in three steps:

1. **Step 1**: Creation of Experience Profiles
2. **Step 2**: Creation of O*NET Job profiles
3. **Step 3**: Experience Profile to O*NET profile matching

We describe two matching procedures:

1. Experience Profile to O*NET Profile matching. This matching uses only the profile characteristic scores.
2. Desired Profile to O*NET Profile matching. This matching uses both the profile characteristic scores, and the characteristic importance levels from the desired profile.

Automated Creation of Experience Profile

The dataset shared with you contains a spreadsheet mapping each experience-related question in each survey to one profile characteristic. In each survey, several (typically three or four) questions map to each profile characteristic defined in the dataset.
Each survey question about user experience with either a URE or a job/internship uses a 7-point scale to indicate the level of user agreement with the affirmative statement of the question. Here, 1 means "strongly disagree" and 7 means "strongly agree". The value 0 is assigned to all non-responses.

Let \( Q = \{Q_1, \ldots, Q_n\} \) be all experience-related questions in some survey \( S \). Let \( C = \{C_1, \ldots, C_k\} \) be the list of profile characteristics used, and let \( m : Q \rightarrow C \) be the mapping of questions to profile characteristics. We also denote as \( m^{-1}(C_i) \) the set

\[
m^{-1}(C_i) = \{Q_j \in Q | m(Q_j) = C_i\}.
\]

Let \( q = (q_1, \ldots, q_n) \) be user responses to questions \( Q_1, \ldots, Q_n \) respectively. We define \( c = (c_1, \ldots, c_k) \), the vector of profile characteristic values as follows:

\[
c_i = \frac{1}{|\{Q_j | Q_j \in m^{-1}(C_i) \land q_j \neq 0\}|} \sum_{j:Q_j \in m^{-1}(C_i),q_j \neq 0} q_j.
\]

That is: to find the value of the profile characteristic \( C_i \) that matches questions \( Q_{i1}, \ldots, Q_{im} \) we first omit all zero values from the list \( q_{i1}, \ldots, q_{im} \) of responses to these questions, and then take the mean of the remaining responses.

**Note:** if the user responds 0 to all questions associated with a specific profile characteristic, the value of the profile characteristics for this particular experience profile is undefined (for practical purposes you can set it to 0 or -1).

### Creation of Desired Profiles

Desired profiles, for each profile characteristic \( C_i \in \text{mathbf{C}} \) have two numbers: \( c_i \): the desired level of this characteristic, and \( w_i \): the desired importance of the characteristic. Both values are specified by the user via direct interaction with the CED-M application.

### Creation of O*NET Job Profiles

The profile of an O*NET Job is constructed using the same set of target profile characteristics as those used to create experience profiles and desired profiles\(^2\).\(^1\)

---

\(^1\)"Experience-related" means that the question is mapped to one of the profile characteristics.

\(^2\)In reality, our current implementation uses only 18 out of the 26 profile characteristics. These are the characteristics on which the matching will be performed. Later versions of the application may implement computing profile values for additional profile characteristics.
The process of creation of O*NET Job Profiles is based on the information provided to you in the spreadsheet titled O-Net Computer Science Template for Survey_5_13_2022.xlsx.

In short, this file (which we refer to as the O*Net Profile Template) contains information on how to determine the value of each profile characteristic for each O*NET job.

The O*NET Profile Template is organized as follows.

- Rows 7 through 48 (note - some rows are empty or contain headers) contain the list of profile characteristics whose values for given O*NET jobs we want to establish. (the spreadsheet calls them "Task Characteristics", their names can be found in the column A of the spreadsheet.

- Columns B through P show different O*NET search facets. The O*NET website allows the users to browse its contents (descriptions of different jobs) by selecting a specific facet of the job.

- Intersections of rows and columns are non-empty, whenever a specific O*NET facet (in a column) has an effect on computing a specific profile characteristic (in a row). The text in the spreadsheet cell is a value of the O*NET search facet, for which the O*NET web site has an ordered list of jobs available. The text serves as an HTML link to the O*NET page presenting an ordered list of O*NET job descriptions (ordered from highest level/importance of the given facet to the job, to the lowest).

Example. The first profile characteristic in the spreadsheet (row 7) is "Work Scheduling Authority". Dr. Bachrach has mapped the notion of "work scheduling authority" to five different facets of O*NET Jobs, as follows:

- Facet: Work Activities - Mental Processes. O*NET list for the facet: "Organizing, Planning, and Prioritizing Work - Developing specific goals and plans to prioritize, organize, and accomplish your work."

- Facet: Work Activities - Mental Processes - cont. O*NET list for the facet: "Scheduling Work and Activities - Scheduling events, programs, and activities, as well as the work of others."

- Facet: Work Context - Interpersonal Relationships. O*NET list for the facet: "Work With Work Group or Team - How important is it to work with others in a group or team in this job?"

- Facet: Work Context - Structural Job Characteristics. O*NET List for the facet: "Structured versus Unstructured Work - To what extent is this job structured for the worker, rather than allowing the worker to determine tasks, priorities, and goals?"

---

3An earlier version of this file was made available to you at the beginning of the course. This file is the newest version, and it is available for download from the course web page.
• Facet: Work Context - Structural Job Characteristics cont.” O*NET list for the facet: ”Time Pressure - how often does this job require the worker to meet strict deadlines?”

What does this all mean? The O*NET Profile Template is Dr. Bachrach’s way of establishing where in O*NET data (on the O*NET web site) one can find information about the specific profile characteristics. In the example above, Dr. Bachrach is conveying the information that the profile characteristic Work Scheduling Autonomy is featured in five different lists available on the O*NET web site.

Each O*NET Profile Template cell that has a link leads to an O*NET landing page for a list of O*NET Job descriptions. The job descriptions contained on the list are furnished with one or more values from the following list: Importance, Level, Context.

NOTE: The O*NET Job lists in two last columns (Column O: ”Work Values” and Column P: ”Work values contd.”) do not have numeric values associated with the jobs. For the purposes of this assignment, ignore these two columns in your computations (i.e., do not visit any lists linked to the cells in these columns).

Importance, Level and Context of a given O*NET job description on the list are integer numbers on the scale from 0 to 100.

The Level value specifies for each O*NET job how complex/prevalent the particular facet is. The Importance value specifies how important this particular facet to the job. (Some aspects of a job may be fairly complex/involved, but not very important/critical to the job itself, for example). The Context value is yet another way to indicate the prevalence of the facet being viewed to the job.

Computing the value of the profile characteristic for O*NET jobs. Given an O*NET job, and a specific profile characteristic (a row in the O*NET Profile Template), the value of the task characteristic for the O*NET job is computed as follows.

• Visit each O*NET URL placed by Dr. Bachrach in the row for the specific profile characteristic.
• Find the O*NET job in question on the list of jobs in the URL provided.
  - If Level attribute is present in the list, collect the Level value associated with the job.
  - If Level attribute is absent, but Context attribute is present, collect the Context value associated with the job.
  - If neither Level nor Context attributes are present, collect the Importance value associated with the job.
• Find the largest collected value among all lists where you located the O*NET job in question successfully.

• Convert the largest value of Level into a value on the 1-7 scale, as follows:

\[ \text{converted} = \left\lceil \frac{\text{Largest Value} \times 7}{100} \right\rceil \]

Assign converted value to the profile characteristic.

Example. Consider the Work Scheduling Authority profile characteristic and the O*NET job Construction Managers. Visiting the five lists (see above) for the five facets associated with the Work Scheduling Autonomy we collect the following information about the Level values from these lists for Construction Managers:

<table>
<thead>
<tr>
<th>List</th>
<th>Value to Collect</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work Activities - Mental Processes</td>
<td>Level</td>
<td>87</td>
</tr>
<tr>
<td>Work Activities - Mental Processes - cont.</td>
<td>Level</td>
<td>87</td>
</tr>
<tr>
<td>Work Context - Interpersonal Relationships</td>
<td>Context</td>
<td>89</td>
</tr>
<tr>
<td>Work Context - Structural Job Characteristics</td>
<td>Context</td>
<td>81</td>
</tr>
<tr>
<td>Work Context - Structural Job Characteristics cont.</td>
<td>Context</td>
<td>75</td>
</tr>
<tr>
<td>MAX VALUE</td>
<td></td>
<td>89</td>
</tr>
</tbody>
</table>

The largest collected value is 89, which we then convert as follows:

\[ \left\lceil \frac{89 \times 7}{100} \right\rceil = \left\lceil 6.23 \right\rceil = 7 \]

Thus, your procedure will assign Construction Managers the value of 7 for the characteristic Work Scheduling Authority.

Bulk Conversion. In practice, rather than going over a list of O*NET jobs one at a time, it is best to visit each URL in the O*NET Profile Template spreadsheet once, collect information about the Level values for all STEM (you will have to filter for that) jobs on the list.

IMPORTANT. STORING INFORMATION ABOUT O*NET JOBS IN CED DATABASE. In order to compensate for a fairly sophisticated procedure for scoring O*NET job profiles, we will simplify the requirements for handling O*NET data in the CED Database. Your CED application (CED-A or CED-M) shall store the following information about an O*NET Job:

• Job name (e.g., "Construction Managers")

• URL of the O*NET landing page

• Short description of the job (the text that shows up at the top of the landing page)

4https://www.onetonline.org/link/summary/11-9021.00
In addition to this, the CED database shall store the job profile associated with each job, computed as described above.

The CED Application (CED-A or CED-M) is NO LONGER REQUIRED to have its own landing page for the O*NET job profile. Instead, your application can simply open a frame, and load the contents of the O*NET landing page URL into it (while providing the navigation controls allowing the user to go back to other CED Application pages).

Experience Profile - O*NET Job Profile Matching

Let $C = \{C_1, \ldots, C_k\}$ be a comprehensive list of all profile characteristics.

An experience profile $s = (s_1, \ldots, s_k)$ is a vector of values (each - an integer in the range 1-7, or an undefined value\(^5\) where $s_i$ is the value of the profile characteristic $C_i$ for the given experience.

A job profile $w = (w_1, \ldots, w_k)$ is a vector of values (each either a value in the range $[1, 7]$ or an undefined value) where $w_i$ is the value of profile characteristic $C_i$ for the given O*NET job.

We can compute the similarity between the experience profile $s$ and the job profile $w$ using any of the traditional similarity metrics. Below, we consider two such metrics: Cosine Similarity and Pearson Correlation.

**Cosine Similarity.** The cosine similarity between $s$ and $w$ is computed as follows:

$$\cos(s, w) = \frac{\sum_{i=1}^{k} s_i \cdot w_i}{\sqrt{\sum_{i=1}^{k} s_i^2} \sqrt{\sum_{i=1}^{k} w_i^2}}$$

**Note:** the values $c_i \cdot w_i$ are only defined for those values of $i$ where both $c_i$ and $w_i$ are not undefined. Similarly, the $\sum_{i=1}^{k} s_i^2$ and $\sum_{i=1}^{k} w_i^2$ skip values of $i$ where either $s_i$ or $w_i$ is undefined\(^6\).

The cosine similarity between two vectors of non-negative numbers ranges between 0 (very dissimilar) to 1 (exactly the same).

**Pearson Correlation.** The Pearson correlation between $s$ and $w$ is defined as follows:

$$\text{pearson}(s, w) = \frac{\sum_{i=1}^{k} (s_i - \mu_s) \cdot (w_i - \mu_w)}{\sqrt{\sum_{i=1}^{k} (s_i - \mu_s)^2 \sum_{i=1}^{k} (w_i - \mu_w)^2}},$$

\(^5\)These may be represented as 0, -1, or in any other way.

\(^6\)If you label undefined values as 0, then you can simply compute the sum above verbatim for the correct effect.
where $\mu_s$ and $\mu_w$ are means values in vectors $s$ and $w$ respectively, for all indexes $i_1, \ldots, i_m$, where both $s_{i_j}$ and $w_{i_j}$ are not undefined.

Pearson correlation ranges from -1 (very dissimilar) to 1 (very similar).

In your implementation, feel free to choose one of these two metrics to use in computing the similarity between an experience profile and an O*NET job.

**Finding the Matches.** Given an experience profile, compute its similarity scores with each O*NET job. Find the 10 best matches, and store information about these 10 matched in the CED database.

**Desired Profile - O*NET Job Profile Matching**

Let $C = \{C_1, \ldots, C_k\}$ be a comprehensive list of all profile characteristics.

A desired profile profile is a pair $(s, v)$, where $s = (s_1, \ldots, s_k)$ is a vector of values (each - an integer in the range 1-7, or an undefined value\(^7\) where $s_i$ is the value of the profile characteristic $C_i$ for the given experience, and $v = (v_1, \ldots, v_k)$ is the vector of importances associated with the profile characteristics (each importance is scored on the scale of 1-7, or is undefined - undefined importances can be imputed to take the value 4).

A job profile $w = (w_1, \ldots, w_k)$ is a vector of values (each either a value in the range $[1, 7]$ or an undefined value) where $w_i$ is the value of profile characteristic $C_i$ for the given O*NET job.

We can use a weighted cosine similarity to represent the match score between a desired profile and an O*NET Job profile:

$$weightedCos(s, w) = \frac{\sum_{i=1}^{k} v_i \cdot s_i \cdot w_i}{\sqrt{\sum_{i=1}^{k} s_i^2} \sqrt{\sum_{i=1}^{k} w_i^2} \sum_{i=1}^{k} v_i}$$

**Note:** the values $v_i \cdot c_i \cdot w_i$ are only defined for those values of $i$ where both $c_i$ and $w_i$ are not undefined. Similarly, the $\sum_{i=1}^{k} s_i^2$ and $\sum_{i=1}^{k} w_i^2$ skip values of $i$ where either $s_i$ or $w_i$ is undefined \(^8\).

**Finding the Matches.** Given a desired profile, compute its similarity scores with each O*NET job. Find the 10 best matches, and store information about these 10 matched in the CED database.

\(^7\)These may be represented as 0, -1, or in any other way.

\(^8\)If you label undefined values as 0, then you can simply compute the sum above verbatim for the correct effect.
Putting it together

Implement the matching procedure in your CED Application (CED-A or CED-M). In doing so, you can use the following design ideas:

- O*NET Job profiles can be computed once during the initial data upload/setup process for the CED database. This code does not need to be an integral part of the CED Application (i.e., you can have a script on the side creating all necessary records and inserting them into the CED database). In general, it is expected that both CED-A and CED-M applications start (i.e., start work with their respective users) only after O*NET Jobs are imported from O*NET into the CED Database, and O*NET Job Profiles are created. (Note: the CED-A use cases include use cases for profile characteristic creation, which might somewhat interfere with this statement. We will discuss a simple way of dealing with this in class.)

- CED-A Application automatically creates an experience profile for each survey response in the database as part of the initial database on-boarding (CED-A application has no means of creating a survey via UI - so all work on experience profile preparation is outsourced to the data upload scripts).

- In CED-M completion of a survey shall trigger an automated creation of an experience profile for the completed survey.

- In both CED-M and CED-A - as soon as an experience profile is created, the application can automatically trigger the matching algorithm, find the 10 best matches and archive them in the CED database for future perusal.

- In CED-A desired profiles are uploaded into the CED database during the initial data upload process (CED-A lacks functionality to create new desired profiles).

- In CED-M desired profiles can be created by the user on the fly.

- As soon a desired profile is added to the CED Database, the application shall trigger the matching algorithm, find the 10 best matches and archive them in the CED database for future perusal.

Submission

Each team shall be prepared to demo the work of the matching implementation on the due day (and/or during the next lab period).