# Universiteit Leiden
# ICT in Business and the Public Sector

## Evaluating the Skills Gap in the Labor Market

| | |
|---|---|
| Name: | Maurits de Groot |
| Student ID: | s1676784 |
| Date: | 26/03/2021 |
| 1st supervisor: | Niels van Weeren |
| 2nd supervisor: | Aske Plaat |

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

# Evaluating the Skills Gap in the Labor Market

Leiden University

Maurits de Groot

March 2021

# Reading Guide

This thesis is the final document to fulfill the graduation requirements for the master *ICT in Business and the Public Sector* offered at Leiden University. Information presented in this thesis is the result of a project supervised by Niels van Weeren and Aske Plaat. For this thesis an internship opportunity is provided by Randstad groep Nederland. Randstad provided resources such as computational capacity, structured information and domain expertise. These resources were used to conduct this research. I would like to thank everybody involved with this project, especially Niels van Weeren, David Graus and Jelle Schutte.

When reading this document, different parts can be identified. First, the problem at hand is explained as well as the rational behind the problem. This part is a good fit for the business minded reader. Multiple sub-problems and use-cases are discussed to provide a complete overview. The second part consists of technical solutions to solve the business problems. This part uses advanced models alongside in depth analyses of the problem. Lastly, the third part uses the findings of the technical solution to answer the business driven questions in the conclusion. The first part is Chapters 1 and 2, the second part consists of Chapters 3 till 6. The final part is Chapter 7 and 8.

# Abstract

The labor market is constantly evolving. Occupations are being changed, created or removed to fit the needs of today's market. In recent years the pace of this change has been accelerated due to factors such as globalization, digitization and the shift to working from home. A number of factors are relevant when selecting employment, e.g. the cultural fit, compensation and degree of freedom provided. This work focuses on the skills required for occupations.

To successfully fulfill an occupation the gap between the skills required and the skills possessed by an individual needs to be as small as possible. Decreasing the skill-gap improves the fit between a job candidate and occupation. This gap occurs in multiple situations, every situation needs to be handled accordingly. When an individual is yet to acquire skills the most relevant skills for the desired occupation need to be learned. If the precise occupation is unclear but the sector of desired work is known, the most relevant skills of that sector need to be learned. The most relevant skills are determined in this work by using a Term Frequency-Inverse Document Frequency (TF-IDF) based model.

In the situation where an individual already has a skillset but no employment, the occupation most related to the skills need to be modeled. Doing so helps to select an occupation that fits with current capabilities. This is done by using the Node2Vec link prediction model. In the last case a job candidate already has a job but is seeking the next job. Here the most efficient transition between jobs need to be mapped based on the jaccard similarity measure. Every situation requires a different approach, this thesis provides fitting techniques to help close the gap.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

In recent years the number of people that change their job is increasing [14], the average duration of a position is shorter [28] and the total working population is growing [29]. This results in a rapidly increasing number of potential job candidates. Candidates enjoy, on average, a higher level of education compared to a number of years ago [2]. This diverse workforce has to face new challenges to secure employment. Due to increasing globalization, the number of possible job candidates per position is higher. As a result of this the labor market is more competitive than it has ever been [3]. Among others, these factors result in a dynamic labor market where changes keep occurring. In the modern labor market it is important to have the rights skills. The increasing amount of digitization has made computer skills more valuable [32]. The COVID-19 pandemic has resulted in a double-disruption effect where technological adoption is accelerated and companies lay off employees [15]. Most aging workers do not posses those technical skills which leads to lower job opportunities for these workers [8]. Not only technical skills are important, having good people skills is becoming increasingly important as well [7]. With the demand of different skills changing over time, having the correct skills for specific occupations is more crucial than it has ever been. Being able to keep up with the latest developments has become more difficult with the accelerated changes in the labor market. The volatility in the labor market results in a change of occupations with new required skills. To find a match between vacancies and job postings, individuals can use external services that match their skills with their desired work. In 2019, employment agencies were responsible for fulfilling $10\%$ of the available jobs in the Netherlands [10]. In recent years the tasks associated with an occupation have become increasingly divergent. Different tasks require different skills to fulfill the task. As a result of this, matching occupations with employment has become skill based, since the desired profile for a given occupation is no longer unambiguous.

## 1.1 Background

The underlying problem which is addressed in this thesis is the unemployment rate, as there is a mismatch between people that are seeking an employment and the employment itself. A population can be divided into two groups: people that are in the labor force and people that are not in the labor force, the latter is addressed as out of the labor force. The group out of the labor force consists of people that are too young to be employed, are retired from the labor force, unwilling to work or unable to work. This leaves the labor force as a group that is willing and able to work. A part of the labor force is working for pay, this part is called the employed part. The rest of the labor force is actively looking for work but currently not employed, this part is called unemployed. The unemployment rate is the fraction of unemployed people from the total labor force, or in a more formal definition:

$$\text{unemployment rate} = \frac{\text{unemployed people}}{\text{total labor force}} \times 100 \tag{1.1}$$

For the purpose of this thesis we consider the lowering of the unemployment rate by moving people from unemployed to employed as beneficial. A lower unemployment rate could have a number of negative aspects which occur mostly on a macroeconomic level [23]. When looking from the perspective of an employment service contributing to a lower unemployment rate is good.

Most of the time unemployment is caused by a mismatch between the requested skills and the skills which an employment candidate possesses. For this thesis we divide unemployed people in four groups based on the following two attributes: (1) whether a candidate supplies the required skills for a given occupation and (2) whether the skills are demanded for a given occupation. A visual representation of the groups can be found in Figure 1.1. The following four groups are formed:

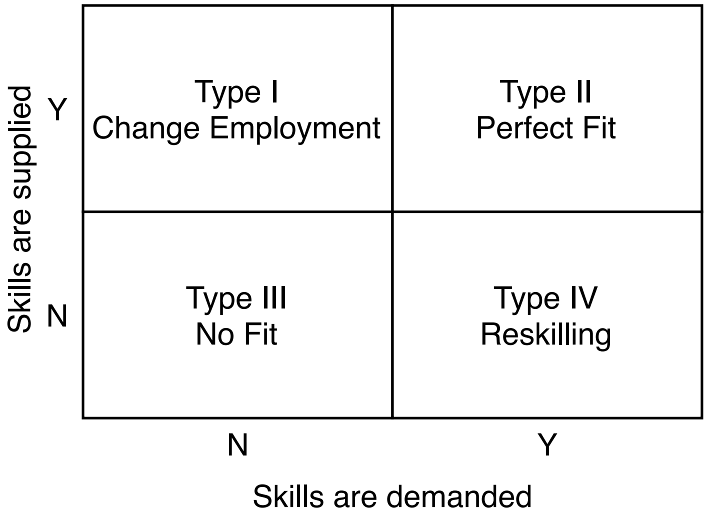


**Figure 1.1:** Matrix with the four types of possible scenarios when matching a candidate with an occupation

I *Supplied skills are not demanded for a given occupation*: There is no demand for the skills which are supplied. The skills are the correct skills for the occupation at hand but the occupation has no demand at the moment. Here a candidate has the right requirements for an occupation but the market for that occupation is saturated.

II *Supplied skills are demanded for a given occupation*: Both the skills that a candidates possesses and the required skills for an occupation are in balance. Here the candidate is a good fit for the function based on the skills.

III *Skills which are not demanded are not supplied*: The skills that a candidate possesses are not the right fit for this occupation and the occupation is not demanding the right skills at the moment. Here, one is unfit for a function that is not hiring.

IV *Demanded skills are not supplied for a given occupation*: A candidate does not have the correct skills for this occupation. Here, the occupation is still hiring for this position.

When the demanded skills for a given job posting are supplied by a job candidate the job candidate is able to do this occupation. If the job candidate is also willing to do the job then the job posting can be fulfilled. A perfect fit between job candidate and demanded skills can be achieved for multiple job postings. State *II* is the most favorable state to be in for job postings which one is willing to do. If one is not willing to work in a given occupation but is in state *II* the job posting will not be fulfilled. One can transition between states in the following ways:

- From *Change Employment* to *Perfect Fit* cannot be achieved for the same job, if no demand is present for a given occupation this demand cannot be created by an individual. However, external events can have an influence on the demand for a given skillset. Switching to an occupation where similar skills are required and demand is present would be the most efficient way to transition.

- From *Reskilling* to *Perfect Fit* can be achieved by learning the required skills. If a skill gap is present learning new skills can decrease or even close this gap. When learning skills it is often harder to learn a complete new skillset compared to extending the current skillset. Learning skills which are related with one's current skills often yields more success.

- Going from state *No Fit* to *Perfect Fit* means that a job candidate wants to have an occupation which the candidate is not equipped for and the function is not available. In this case the candidate could look at occupations that are related to the desired one but closer to the current skillset of the candidate.

## 1.2 Problem Statement

To be able to lower the employment rate people need to be reskilled or switch to jobs with are related to the jobs that have become unavailable. Doing so is equivalent to transitioning to the *Perfect Fit* state in the model introduced in Figure 1.1. To facilitate this, one needs to know which skills are relevant for which occupations. A notion of similarity is required to define which skills or occupations are related. Here, the similarity between skills and other skills is needed to discover which skills could be learned more easily when possessing other skills. It is relevant to know which occupations are alike to help people find a new job which is similar to their old job in terms of skill requirements. Knowing which skills are related with a number of occupations tells which skills need to be learned when an occupation is desired. In a rapidly globalizing economy this information needs to be language independent. It needs to be updated in a robust and accurate way to ensure that it cannot get outdated. Additionally, it needs to be adaptive to new situations such are the creation and deletion of skills or occupations.

Having this information enables a number of use cases such as:

- Starters Matching: finding the best occupation to place a starter given a set of skills.

- Talent Search: knowing which skills are most important for a job posting that needs to be filled.
- Reskilling: Teach new skills to an individual to change the capabilities of that individual to fit the market better.
- Career Pathing: helping individuals to future-proof their careers by suggesting relevant skills to learn to transition from one occupation to the next.
- Skill Relevance: determining which skill is the most relevant skill for any given occupation.
- Market Insights: knowing the latest status of the labor market is essential when making strategic decisions.

The use cases listed above are reflected in the research questions in Section 1.4.

## 1.3 Scope & Context

What a job or occupation is could be widely interpreted depending on different viewpoints. For the purpose of this thesis the definitions of the International Standard Classification of Occupations (ISCO) [19] will be used. This is the industry standard for occupation classification. When ISCO is mentioned in this document, it refers to the version ISCO-08. The ISCO definitions are as follows [20]:

**Job:** "a set of tasks and duties performed, or meant to be performed, by one person, including for an employer or in self-employment"

**Occupation:** "set of jobs whose main tasks and duties are characterized by a high degree of similarity"

The ISCO is developed as an taxonomy to classify occupational groups with four granularity levels across ten different major groups. The mayor groups defined by ISCO can be found in Table 1.1. The four ISCO groups are ordered as a hierarchical classification scheme, such a scheme is also known as a taxonomy. In this taxonomy an occupational group has multiple levels. Here, a computer programmer is defined by the level 4 ISCO code: 2132. This means that this occupation is part of the level 1 group professionals (ISCO-code "2"), the level 2 group computing, engineering and science professionals (ISCO-code "21") and the level 3 group computing professionals (ISCO-code "213").

Within the major ISCO groups a different number of sub groups are defined. The sub groups are unevenly distributed, this becomes apparent when looking at the unit group (level 4). Here, the smallest group consists of $3$ objects. The largest major groups have a total of $92$ objects on the unit group level. This means that in the major group Professionals more distinct occupations are present compared to the Armed forces occupations. An overview of the number of categories for every ISCO level can be found in Table 1.2.

By following the ISCO we have a definition for what an occupation is. Since the ISCO is a taxonomy it does not include skills. When matching occupations with skills both need to be

| Group Number | Major Group Name |
|---|---|
| 1 | Managers |
| 2 | Professional |
| 3 | Technicians and associate professionals |
| 4 | Clerical support workers |
| 5 | Service and sales workers |
| 6 | Skilled agricultural, forestry and fishery workers |
| 7 | Craft and related trades workers |
| 8 | Plant and machine operators, and assemblers |
| 9 | Elementary occupations |
| 10 | Armed forces occupations |

**Table 1.1:** The 10 major job groups of the ISCO-08

| ISCO | Major Groups (level 1) | Sub-major Groups (level 2) | Minor Groups (level 3) | Unit Groups (level 4) |
|---|---|---|---|---|
| Group 1 | 1 | 4 | 11 | 31 |
| Group 2 | 1 | 6 | 27 | 92 |
| Group 3 | 1 | 5 | 20 | 84 |
| Group 4 | 1 | 4 | 8 | 29 |
| Group 5 | 1 | 4 | 13 | 40 |
| Group 6 | 1 | 3 | 9 | 18 |
| Group 7 | 1 | 5 | 14 | 66 |
| Group 8 | 1 | 3 | 14 | 40 |
| Group 9 | 1 | 6 | 11 | 33 |
| Group 10 | 1 | 3 | 3 | 3 |
| Total | 10 | 43 | 130 | 436 |

**Table 1.2:** Number of categories within the 10 major ISCO-08 groups on levels 1 - 4

defined. The European Skills, Competences, Qualifications and Occupations (ESCO) will be used for the definition of skills [13].

**Skill:** "the ability to apply knowledge and use know-how to complete tasks and solve problems"

The ESCO is an ontology that contains 2942 occupations and 13485 skills in 27 languages. This ontology links different concepts, such as a skills or occupations, with each other. Here, a network is created that relates the skills that could occur for a given occupation with that occupation. Multiple occupations can match a level 4 ISCO group. In Figure 1.2 the link between ISCO and ESCO is illustrated. Note that in this figure ESCO occupations are displayed. These occupations have skills linked to them as well. By using both ISCO and ESCO we get well defined definitions which are common in both industry and academia. The method that is demonstrated could also be used for different combinations of taxonomies and ontologies. As long as a notion of skills and occupations is present. Combining formal standards will provide a theoretical overview of the current labor market. This understanding

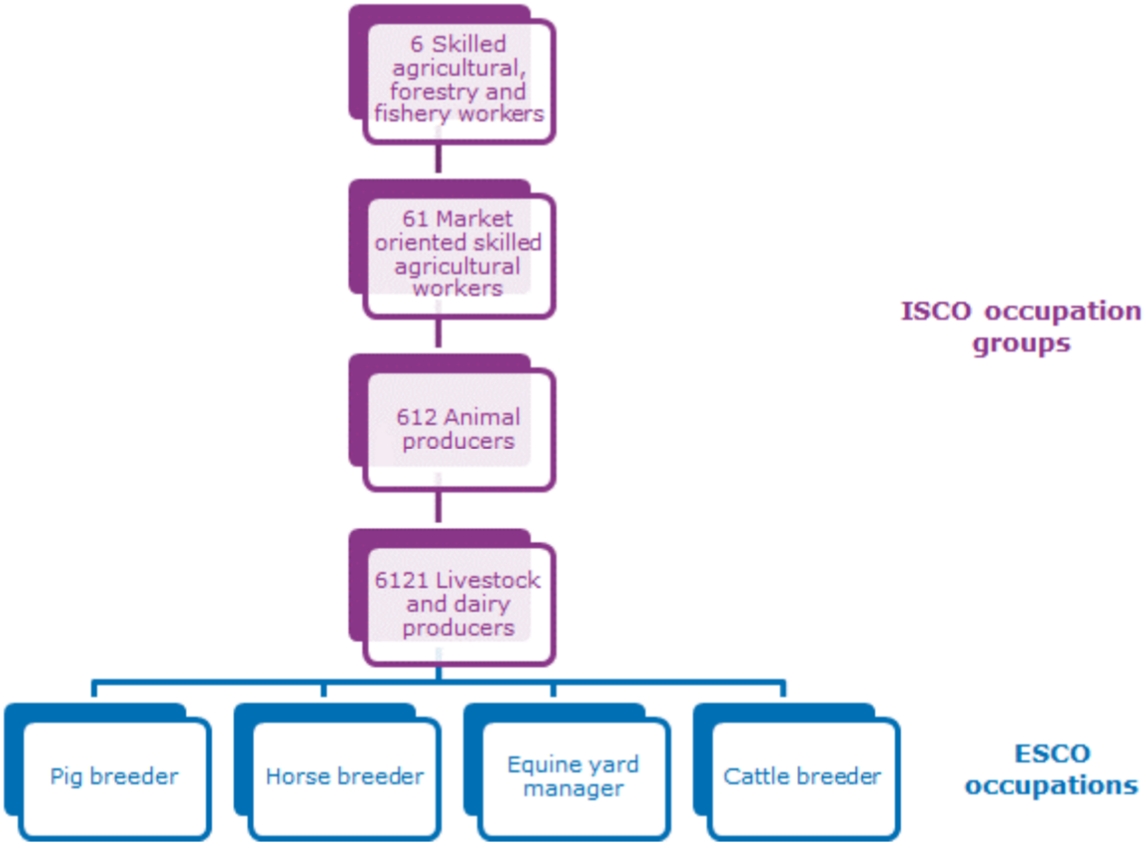

6 Skilled agricultural, forestry and fishery workers
61 Market oriented skilled agricultural workers
612 Animal producers
6121 Livestock and dairy producers
ISCO occupation groups
Pig breeder
Horse breeder
Equine yard manager
Cattle breeder
ESCO occupations

**Figure 1.2:** The structure of the occupations pillar [13]

is limited to the creation date and therefor not up-to-date by definition. To ensure that the latest status of the labor market is represented the ISCO and ESCO standards will be enriched with vacancy data. This data is used to extract the skills which are currently requested for any given occupation. To do so, the job postings used to extract the skills from need to be labeled according to the taxonomy in use, which is ISCO in our case. The output of this exercise should be skills with a corresponding ESCO skill code.

## 1.4 Research Question

Skill frameworks are a volatile area. Most ontologies are made at a fixed point it time, this limits the usability of them when time passes. Using an expired reference point to match jobs results in sub-optimal results. To be able to match skills and occupations in an optimal way the data used for decision making should be up-to-date. When using such a framework it needs to be able to self-correct when the situation on the labor market changes. It should be adaptive to new skills and occupations. The aim of this thesis is to create a model that is able to extend the current ESCO ontology. To reflect the latest labor market developments real world job postings will be used. This model should be able to answer the following question:

**RQ:** *How can the perfect fit between a job candidate and occupation be found based on skills?*

To answer this question a number of sub-questions are defined, each sub-question reflects the need of the business represented by the use cases from Section 1.2.

**Q1:** *How can we model the notion of relatedness between a skill and an occupation?*

In the use-case of *talent search* a new job posting needs to be posted. Knowing what skills are relevant for an occupation is key when connecting talent with this job. If the posting itself does not provide accurate requirements the posting will not be filled successful.

**Q2:** *Given two occupations, what is the most efficient transition based on skill similarity?*

Here a transition is switching between occupations by ensuring that the required skills for the new occupation are collected. For the use-case *career pathing* an individual wants to future proof their career when transitioning from one job to the next. To grow from a starter job to a more advanced, well desired one, the right path needs to be followed. Knowing this path and the most efficient way to follow it enables one to make this step. If an individual does not have the required skills for a job the individual needs to enlarge their skillset by learning new skill, also referred to as *reskilling*.

**Q3:** *How can we determine what skills are the most relevant in any depth in the ISCO taxonomy?*

Using this information the *skill relevance* use-case can be solved. Knowing the most relevant skills for any set of occupations provides *market insights* as well.

By answering question 1 we are able to link any skill to any other skill and determine a notion of closeness to discover which skills are most related with each other. We will also be able to do the same for any combination of skills and occupations or occupations and occupations. Being able to do so helps with creating recommendations for transitions between occupations. The second question creates opportunities to improve career pathing, it helps individuals to keep advancing their careers. The answer to the last question will provide insights in the labor market, here the market could be monitored at multiple granularity levels.

## 1.5 Methodology & Structure

In this thesis the skills in the ISCO taxonomy and occupations from the ESCO ontology will be combined. This combination is used as a formal definition of the inter-cooperation between skills and occupations. This definition provides the theoretical structure which can be build upon. Having this structured information is useful, however it is not a good indication of the current labor market. The labor market is constantly changing, new skills and occupations will arise and older ones will vanish due to the absent of demand. As a proxy for the current situation in the labor market job postings are used. From these job posting, candidate skills will be extracted to model the real demand.

The candidate skills will be matched with the skills as defined in the ESCO ontology to create a knowledge graph that is able to depict the latest status of the labor market. This process as described above is displayed in Figure 1.3. The graph consists of both structured data originating from the ISCO and ESCO frameworks and noisy, unstructured data originating from the job postings. In combining information will be lost, the completeness of the graph

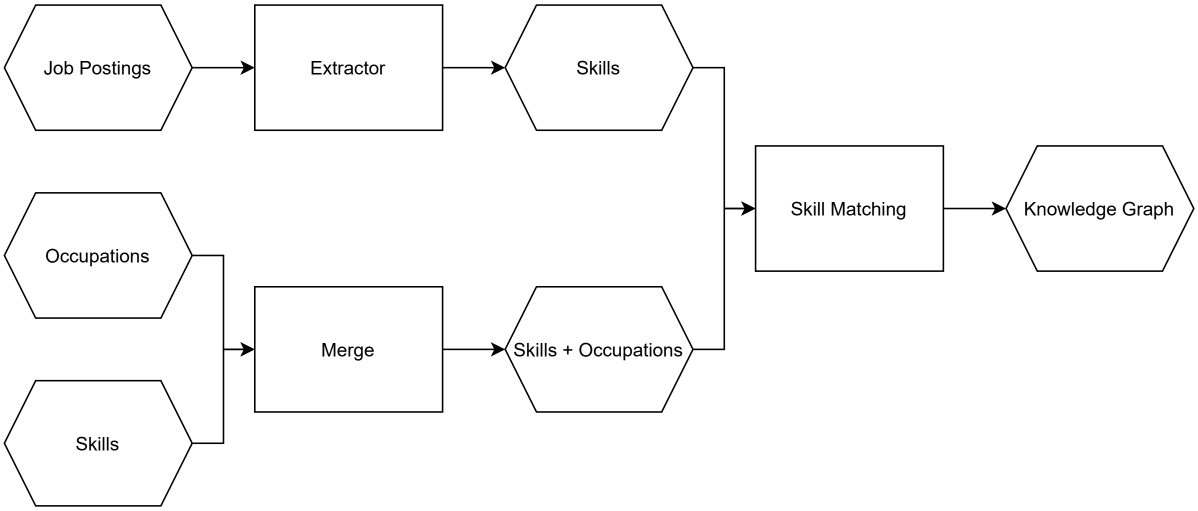


**Figure 1.3:** Creation of the knowledge graph

will become depended on the inclusiveness of the skills extracted from the job postings. If an occupation is not represented (properly) in the job positing, it will not be in the final graph.

The structure of the thesis is as follows. First relevant concepts are explained in Chapter 2. The process of creating the knowledge graph can be found in Chapter 3, here the steps taken to collect, clean and process the data are explained in-depth. Once the graph is constructed the first research question is answered in Chapter 4 by applying link prediction techniques. Using link prediction enables to model the relatedness of skills and occupations. After that set based similarity is used in Chapter 5 to model the notion of closeness between skills. By having a weighted distance between two skills or two occupations, the most efficient transition between any pair of occupations can be calculated. The last research question will be answered in Chapter 6, where relative relevance is determined by taking the Term Frequency Inversed Document Frequency (TF-IDF) score of every skill in every ISCO level for every occupation. This results in insights regarding the relevance of skills. After that in Section 7 the research will be concluded. Lastly in Chapter 8 the discussion and suggestions for future research can be found.

# Chapter 2

# Literature Review

To understand the concepts in this thesis some background knowledge is required. In this section relevant literature is presented to elaborate on the used concepts. This section is divided into a number of sections, each section explains a different concept.

## 2.1 Knowledge Graph

When describing a connection between an occupation and a skill, a structure needs to be used to save this information. For the purposes of this research we opted to use a graph structure to save this information. The advantage of using a graph, compared to a more traditional table based approach, is that related skills can be found directly by looking at part of the graph. An example of a table based representation can be found in Table 2.1. The corresponding graph is displayed in Figure 2.1. Note that both representations contain the same information, however, finding shared occupations for any given skill is computational less expensive for the graph representation. A fraction of the information needs to be inspected to ensure that all neighbors of a given skill are found.

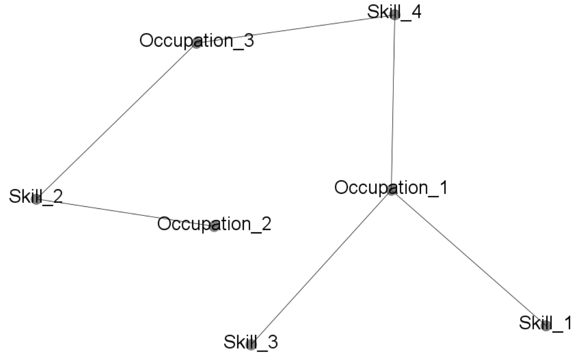


**Figure 2.1:** A graph based representation

| Occupation | Skill |
|---|---|
| Occupation_1 | Skill_1 |
| Occupation_1 | Skill_3 |
| Occupation_1 | Skill_4 |
| Occupation_2 | Skill_2 |
| Occupation_3 | Skill_2 |
| Occupation_3 | Skill_4 |

**Table 2.1:** A table based representation

A graph is defined by the vertex set of a graph and the edge set of a graph [39]. For a graph the notation $G$ is used. The edges (or links) of graph G are denoted by $E(G)$. The vertex set of graph G is denoted by $V(G)$, the terms vertex and node are used interchangeably. When nodes are directly connected the term neighborhood is used. The neighborhood of vertex $v$ is denoted by $N(v)$. The set of adjacent vertices of $v$ are:

$$N(v) = \{x \in V | vx \in E\} \tag{2.1}$$

meaning that $x$ is a neighbor of the vertex $v$ if $x$ is within the vertex set and that in the edge set the relation $xv$ is present. The collection of nodes where these conditions hold are the set of neighbors of $v$. In this thesis a vertex could be an occupation or a skill, both concepts are used within the same graph. The relationship between those entities vary depending on the type of entity. When a graph-structured data model is used to integrate data from different entities the graph is referred to as a knowledge graph [31]. The idea of linking data from multiple sources to create a graph which could provide novel insights is not new. This has been done since the 1980s [27] and is still in use in more recent studies [5]. The quality of the resulting knowledge graph is measured by using the following measurements [31]:

**Coverage**: *The graph contains information about each and every entity in the universe*

**Correctness**: *The information in the graph is true 100% of the time*

When creating a knowledge graph the resulting graph will never be perfect [6]. Generally there is a trade-off between coverage and correctness [40].

## 2.2 Similarity Measurement

To measure the distance between two vertices a similarity measurement is used. A score between 0 and 1 is given to express how similar the entities are. If two identical entities are measured the similarity score will be 1, since any entity is 100% the same as itself. If two entities are completely distinct a similarity score of 0 will be assigned. When a similarity score is known, the distance between two entities can be calculated in the following way.

$$\text{distance} = 1 - \text{similarity-score} \tag{2.2}$$

If two entities are similar the distance between the entities is low. A similarity score is determined by a similarity measurement. Different measurements can be used. One of the most well known similarity measurement is the jaccard similarity [21].

$$J(A, B) = \frac{A \cap B}{A \cup B} \tag{2.3}$$

Here, $A$ and $B$ are sets of objects. The ∩ symbol denotes the overlap in sets. the ∪ symbol denotes the union of the sets. Both sets $A$ and $B$ need to be finite. For calculating the distance between two occupations we can define an occupation as the set of linked skills. Following the example in 2.1 we could define A as occupation_1 and B as occupation_3. This would result in the following sets:

$$A = \{skill\_1, skill\_3, skill\_4\}$$

$$B = \{skill\_2, skill\_4\}$$

The intersect (∩) consists here of 1 element since skill_4 is both present in $A$ and $B$. The union (∪) consists of 4 elements since the total number of distinct elements is 4. This would result in a jaccard similarity of $0.25$ and a distance between occupation_1 and occupation_3 of of $0.75$.

## 2.3 Link Prediction

Within graphs, link prediction is the problem of predicting the existence of an edge [25]. If two nodes are connected by an edge the prediction should be positive, if nodes are not connected the prediction should be negative. Predicting a link can be done in a number of ways. The main ways of link prediction are one of the following [1]:

1. Topology-based
2. Content-based
3. Mixed methods

Topology-based approaches operate under the assumption that nodes with a similar structure, i.e. shared neighbors, have a higher probability of being linked. Content-based methods use the node attributes to predict links. Mixed methods combine topological and content based methods. This often results in more complex methods which yield better results in terms of prediction accuracy. The downside of the increased complexity is expressed in higher computational costs.

### 2.3.1 Node2Vec

One of the more well established link prediction methods is Node2Vec [18]. This method is part of the mixed methods. This algorithm is based on the Word2Vec [26] algorithm which takes a number of sentences as input and uses a skip-gram model to create embeddings. This algorithm has proven successful in numerous situations. To use the same underlying logic in the Node2Vec model a graph needs to be represented as a sentence. For a computer a sentence is an array of tokens. In a linguistic sentence the tokens are represented by words while in the tokens in a graph are represented by nodes. In Figure 2.2 this comparison is visualized. To create an array of tokens from a graph the Node2Vec algorithm uses random walks. A random walk selects a node in the graph as a start and moves from this node to one of the adjacent nodes. For the random walks a number of basic parameters can be selected:

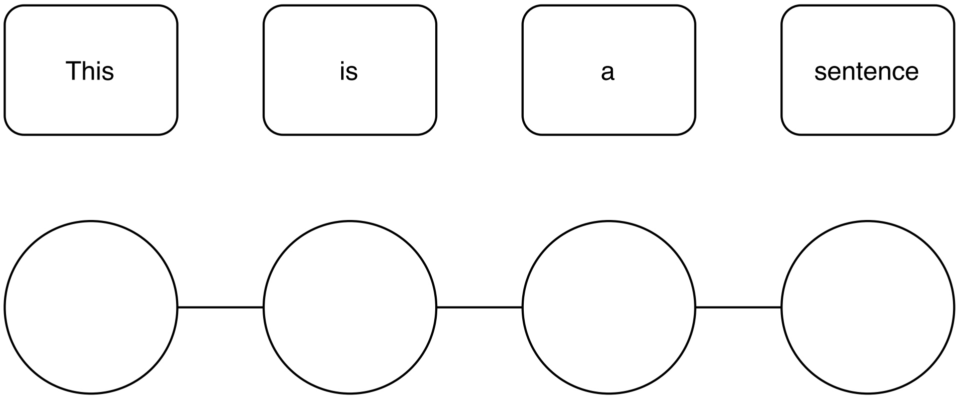


**Figure 2.2:** Two arrays of tokens where the upper tokens are represented by words in a sentence and the lower tokens are represented by nodes in a graph

- Output Dimension: The dimension of the created embedding.

- Number of Walks: The number of walks in the graph
- Walk Length: The length of each walk in the graph

Beside those parameters the Node2Vec algorithm uses two additional hyperparameters. The return parameter ($p$) and the in-out parameter ($q$). In Figure 2.3 part of a graph is shown. In this graph a random walk just appeared from $t$ to $v$. From node $v$ points $t$, $x_1$, $x_2$ and $x_3$ are accessible. The probability of a transition is denoted by $\alpha$. To influence the search bias $p$ and $q$ can be set. The return parameter ($p$) controls the chance that we travel backwards from the node where we are to the node where we have just been. A high value for $p$ creates a lower chance for this happening. The in-out parameter ($q$) allows to differentiate between inward and outward nodes. If a value of $q > i$ is selected the random walk approximates the behavior of a breadth-first search [24]. If a value of $q < 1$ is selected the random walk will approximate the behavior of a depth-first search [36].

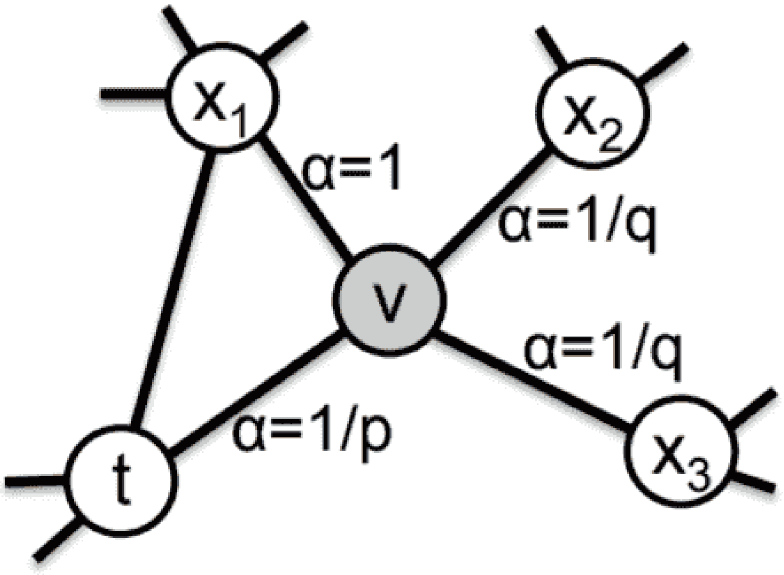


**Figure 2.3:** Illustration of the random walk procedure in node2vec. The walk just transitioned from $t$ to $v$ and is now evaluating its next step out of node $v$. Edge labels indicate search biases $\alpha$ [18]

# Chapter 3

# Data Preparation

To create the knowledge graph of this research a number of steps are required. Part of the process is combining the ISCO and ESCO standards to an ontology. This is a trivial task since the ISCO level 4 groups has a direct match with ESCO objects. Combining those is done via a simple match on the provided identifiers. The more difficult task in creating the knowledge graph is the extraction of skills from the vacancy data. The dataset that is used for the vacancies is described in Section 3.2. The task of extracting skills from the vacancies is described in Section 3.2. When the skills are extracted they need to be matched with existing skills from the ESCO framework. Doing so has two main advantages: (1) following a predefined definition of a skill and (2) eliminating noise from the data. Techniques used for matching are described in Section 3.3.

## 3.1 Vacancy data

The vacancy data used in this research originates from the company Jobdigger [17]. This company scrapes a large number of job postings from different websites. Every job posting is labeled with a level 4 ISCO code. Using this code the sector of the job can be deducted. From all the scraped vacancies a sample is taken of $600.000$ distinct job postings. This sample consists of an equal distribution in ISCO level 1 job postings. From each major ISCO group a random sample is taken to represent job demand in that group. The vacancies in this dataset are Dutch vacancies. These postings are optimized by an internal team from Randstad where low quality job postings are discarded, this was done before taking the sample and does not influence the sample size.

## 3.2 Vacancy Parser

Extracting skills from vacancy texts is a difficult task by itself. One could write a complete thesis on this subject, therefore a commercial of the shelf solution is used for this step. To do this task two industry standard solutions are tested, Burning Glass LENS [16] and Textkernel Extract [37]. These parsers are compared by taking a random set of vacancies and using the parsers to extract skills. The random set that was used to test the parsers was the same for both parsers. From that random set a smaller subset was used to review the performance by hand, based on this the Textkernel parser was selected. This selection process was done in cooperation with an internal team at Randstad. Note that these parsers are developing, at a different moment in time a different parser might be selected.

## 3.3 Skill Matching

To match skills with each other we need to define what it means if a skill matches. This can be done in a number of ways. A strict matching could be enforced where every character of two words need to be equal. For example the skill “administration” is the same as “administration”. Using strict matching ensures that every skill is matched to the desired counterpart, however, matching the skill “administration” to “Administration” (note the capital a) will not result in a match. Both skills clearly have the same meaning so this needs to be corrected. To do so, skills can be normalized before comparing. This ensures that skills like “administration”, “administration ”, “administration1” or “administration)” all result in “administration”. Taking the normalization steps improves our result. Using strict matching minimizes the amount of false matches, but it also decreases the total amount of matches. The skills “maintaining administration” and “administration” have the same meaning but will not result in a match following strict matching. To match these skills we need to find a ways for skills that are alike but not the same to match as well. When doing so we want to preserve the structure of the skills. A match between the individual components of a skill can be made by using n-grams. Here a word get will be represented as a collection of consecutive characters of the word with a fixed length. The value for $n = 5$ has been chosen based on the average length of the skills. This length is calculated after the normalization steps. After normalization the average length is $8$. By taking a $n$ value of $5$ the n-grams are long enough to contain the structure of the skills but not so long that a small number of representations will be created. A fixed length of 5 does not provide a result for skills with a lower number of characters than 5. To solve this the $n$ value is determent in the following way:

$$n = \begin{cases} v & \text{if the length of the skill } (v) \text{ is lower than } 5 \text{ characters} \\ 5 & \text{otherwise} \end{cases} \tag{3.1}$$

Comparing two n-grams should yield a similarity score. For this score jaccard similarity is used as described in Section 2.2. The resulting score can be used to determine how good a match is. If two equal skills are matched the resulting similarity index is 1 and the distance is 0. Lower jaccard distances provide fewer but higher quality matches. To select the best threshold for the jaccard distance a threshold of 0.5 was taken initially. For this threshold 100 random pairs of ESCO skills and vacancy skills are selected. If almost all pairs are reasonable in a manual review the threshold is too harsh and can be set higher. After inspecting the thresholds 0.5, 0.6, 0.7, 0.65, 0.66 and 0.67 a threshold of 0.66 was selected. Here the trade-off between available skills and correctness was most favorable.

The process of matching skills consists out of a number of steps. First, the skill is normalized. After that the n-gram is taken with a $n$ value based on the character length of the skill. The n-gram of every skill that originates from the job postings is then compared to the n-gram of every skill in the ESCO standard based on jaccard distance. An overview of this process can be found in Figure 3.1. In this figure only 1 candidate skill is shown alongside 1 ESCO skill. This process repeats for every candidate skill and every ESCO skill. In total $39.758.827$ candidate skills are compared against $13.485$ ESCO skills, this results in $536.147.782.095$

comparisons. The constructed graph has a total of $1220$ nodes, these nodes consist of both skills $(983)$ and occupations $(237)$. Every node consists of an occupation from the ISCO taxonomy or a skill from the ESCO ontology. These nodes are connected by $3910$ edges. The graph has an average degree of $6.4$, meaning that on average every node has $6.4$ connecting edges. Looking at the combined ontology of ISCO and ESCO this seems like a relative small amount, a complete graph would consists of $13.485$ skills and $436$ occupations. One could argue that, due to its small size, the graph would not be representative. However, the nodes in the graph originate from the demand in the market, represented by job postings. This would make the graph not as representative for the ISCO and ESCO frameworks but more representative for the labor market.

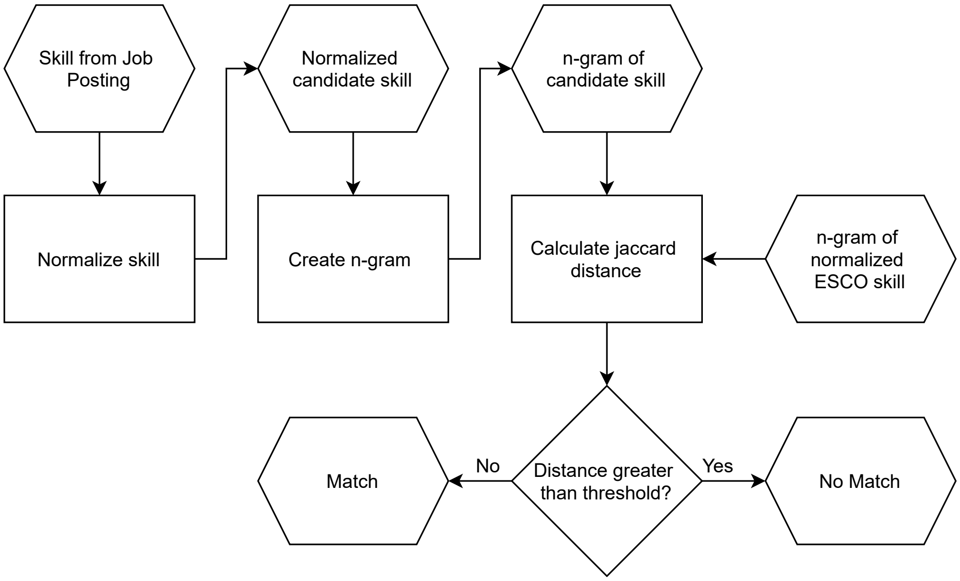

Skill from Job Posting
Normalized candidate skill
n-gram of candidate skill
Normalize skill
Create n-gram
Calculate jaccard distance
n-gram of normalized ESCO skill
Match
No
Distance greater than threshold?
Yes
No Match

**Figure 3.1:** Overview of skill matching process

# Chapter 4

# Modeling the Relations between Skills and Occupations

When searching for talent to fulfill an occupation the right skills are required, but knowing the right skills is not always trivial. In a rapidly changing labor market one cannot be an expert in every field. When decisions are solely taken by human actors biases are introduced. A more objective way of modeling the skills required for an occupation is by using a machine assisted approach. In this section a method is proposed where the relatedness between skills and occupations is quantified using link prediction techniques. This allows to differ from the binary related/unrelated to a intermediate score. Producing this score helps when ranking skills for an occupation.

To model the relation between a skill and an occupation, different techniques can be used. Within the constructed knowledge graph skills and occupations are directly linked. Due to this structure no skills are directly connected with other skills and no occupations are linked with other occupations. Within graph science the prediction of links between two nodes is called *Link prediction*, this subject is examined in Section 2.3. In this chapter different link prediction techniques are compared. When a good link prediction algorithm is selected the unseen links between skills and occupations can be modeled based on the existing structure of the knowledge graph. When such algorithms are trained it is key that the algorithms learns the general structure of the data and not merely store the given information in a different form. To ensure that algorithms represent the given information in a generalized manner the data that is used to train the algorithm is separated from the data that is used to evaluate the performance of it. This is commonly referred to as a training-test split [22].

Most advanced algorithms provide different (hyper)parameters. To select the most optimal parameters a different set of examples is hold back, this is called the validation set. When the data is divided into three parts (training, testing and validation) this is referred to as the train-test-validation split [22]. For this research 55% of the data is randomly sampled to train the algorithms on. The random seed is fixed to ensure that every algorithm has an the same data in the training, testing and validation sets. The testing set consists of 30% of the total dataset, this leaves 15% for the validation set. This distribution is selected such that the number of testing examples is sufficient to depict accurate performance. The validation set is just large enough to have a consistent indicator of performance. For every (positive) example in the training, test and validation set a negative example is added as well. The positive examples are pairs of nodes (edges) that consists in the graph. Negative examples are pairs of nodes that are not present in the knowledge graph, these edges are randomly generated with the following restrictions.

- Random edges in the test, train or validation set do not occur in a different set as well.
- Two identical nodes are not allowed to be a random edge.
- A random edge consists of a skill on one side and an occupation on the other.

To compare the performance the ratio positive to negative examples is 1-to-1, this means that an equal number of positive examples and negative examples are generated. An overview of the amount of edges in every set can be found in Table 4.1.

| | Positive | Negative |
|---|---|---|
| Training edges | 2151 | 2151 |
| Validation edges | 586 | 586 |
| Test edges | 1173 | 1173 |
| Total | 3910 | 3910 |

**Table 4.1:** Number of positive and negative edges with a training (55%), validation (15%), test (30%) split

## 4.1 Preferential Attachment

The first link prediction method is preferential attachment [25]. This method takes a set of nodes, i.e. node $v$ and node $u$, and calculates a closeness between the nodes. This is computed with the following formula:

$$|\Gamma(u)| \times |\Gamma(v)| \tag{4.1}$$

where $\Gamma(u)$ denotes the neighbors of $u$.

Note that the nodes $u$ and $v$ do not need to be neighbors of each other. A higher score correspond here with a greater chance that the nodes are connected. The intuition behind this is that if both nodes have a high amount of neighbors the nodes might function as a hub. Most graphs have the property that hubs have an higher chance to be connected. To create a link prediction for this method the calculated values are saved in a large matrix where every node is represented as both a row-value and a column value. At the intersect of two nodes the preferential attachment value is stored. Note that this matrix is symmetric since the value for row $u$ and column $v$ is equal to the value at row $v$ and column $u$. The matrix is normalized by dividing every value in the matrix by the maximum value found in the matrix. This ensure that values are always between 0 and 1. Resulting preferential attachment values are rounded relative to the average. The resulting value is treated as the chance that the corresponding nodes are related.

## 4.2 Node2Vec

The second link prediction method used is the node2vec algorithm as described in Section 2.3.1. This algorithm can have a number of configurations. For this paper the following parameters are used:

- dimensions = 1024
- walk length = 4
- number of walks = 2500
- $p$ (return parameter) = 1
- $q$ (in-out parameter) = 1

These parameters are selected by conducting a grid search on a large number of possible combinations of parameters. The performance for this parameter selection is tested on the validation set. When testing link prediction algorithms a measurement is required to distinguish between different levels of performance. To do so a number of known edges are tested in the test set. Here, an edge could be positive, meaning that this edge is part of the knowledge grape, or negative, meaning that this edge is not part of the graph. The task of the link prediction algorithm is to predict which edges are part of the graph (positive) and what edges are not (negative).

## 4.3 Evaluation

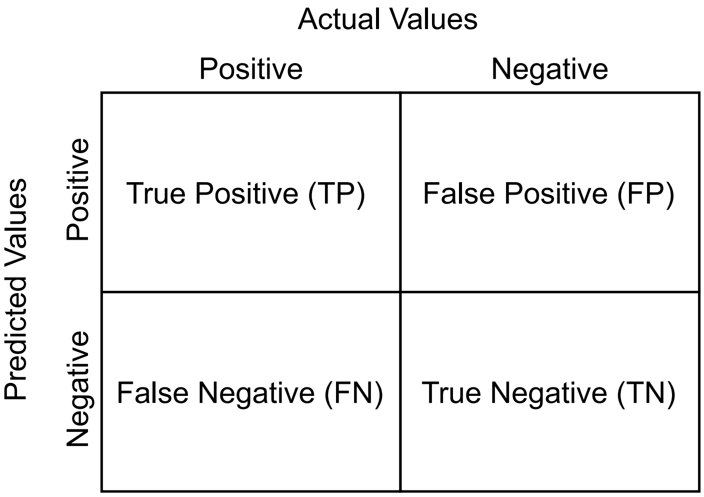
When an algorithm predicts that an edge is part of the graph and the edge is indeed in the graph this is called a *true positive* (tp). If the prediction states that an edge is in the graph but it is not represented in the graph a *false positive* (fp) occurs. The same logic holds for negative examples. If a negative example is predicted correctly the term *true negative* (tn) is used. If a negative relation is predicted where actually a positive link exists a *false negative* (fn) occurs. An graphical representation of this can be found in the confusion matrix displayed in Figure 4.1. To compare different algorithms a number of metrics can be chosen. Two of the most commonly used metrics are precision and recall. The precision [34] metric summarizes the percentage of examples that are correctly classified as positive relative to the total amount of positively classified predictions. This is formally defined as:

$$\text{precision} = \frac{tp}{tp + fp} \tag{4.2}$$

The recall [34] metric states how well the positive class is predicted. This is done in the following way:

$$\text{recall} = \frac{tp}{tp + fn} \tag{4.3}$$

The precision and recall matrics address different concerns, when both are relevant the metrics can be combined in the F1-score [34] (also called f-score). This metric provides the best of both worlds. If the dataset is unbalanced (the amount of positive examples is not equal to



| | | Actual Values | |
|---|---|---|---|
| | | Positive | Negative |
| Predicted Values | Positive | True Positive (TP) | False Positive (FP) |
| | Negative | False Negative (FN) | True Negative (TN) |

**Figure 4.1:** Confusion matrix

the amount of negative examples) the F1-score is commonly used. The formal definition of the F1-score is the following:

$$\text{F1-Score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{4.4}$$

By using the F1-score to evaluate the link prediction algorithms we can see which algorithm is most accurate in predicting the links. When comparing these algorithms we need to take into account that node2vec is a stochastic algorithm due to the randomness of the random walk. As a result of this the outcome of the node2vec algorithm is subject to slight variations. To encounter for this the reported scores are averaged over $100$ runs. The student t-test [35] is used to verify the significance of the findings at the alpha level of $0.01$. Table 4.2 shows the performance of both node2vec and preferential attachment. If an equal number of positive and negative edges are in the test set the preferential attachment method outperforms the more complex node2vec algorithm with an f1-score for the positive class of $0.78$ against $0.65$. This score is taken as a measurement for performance since we want to accurately predict if a link is present. In most realistic situations the possibilities of linking to a positive edge and a negative edge are not 1-to-1. To simulate real world performance this ratio of negative edges / positive edges should reflect more realistic proportions. To do so we calculate the F1-score at multiple ratios, here a ratio of 7 means that 7 times more negative edges are present compared to the number of positive edges. In Figure 4.2 the F1-scores are displayed

| | class | precision | recall | f1-score | support |
|---|---|---|---|---|---|
| Preferential Attachment | 0.0 | **0.83** | 0.64 | 0.72 | 1173 |
| | 1.0 | 0.71 | **0.87** | **0.78** | 1173 |
| Node2Vec | 0.0 | 0.66 | **0.90** | **0.76** | 1173 |
| | 1.0 | **0.84** | 0.53 | 0.65 | 1173 |

**Table 4.2:** Precision, recall and F1-scores of multiple link prediction algorithms with an equal number of positive and negative edges used for training

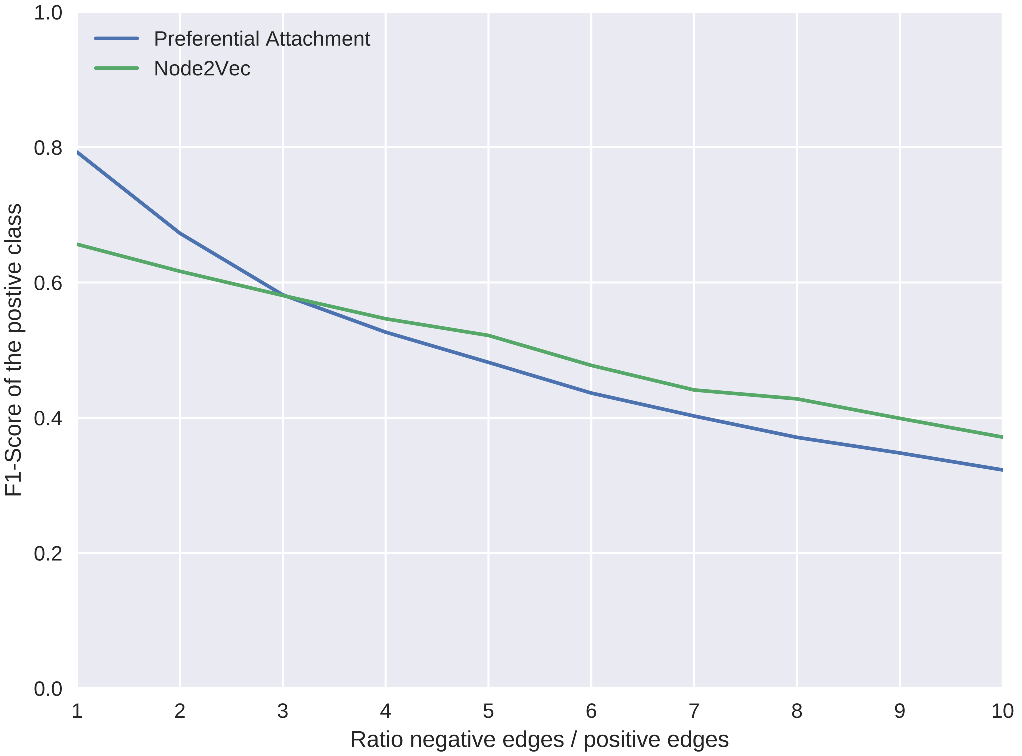


**Figure 4.2:** Comparison of Node2Vec and Preferential Attachment for different ratio's negative edges / positive edges

for multiple ratios. Looking at this figure we see that if the ratio negative edges / positive edges is 3 the performance of node2vec is on par with the preferential attachment method. When the ratio increase node2vec algorithm start to outperforms the preferential attachment method. From these experiments we can conclude that the node2vec algorithm is better suited for most real world situations.

Now that it has been established that the node2vec algorithm is the best fitting algorithm for our dataset we can use this algorithm to predict the relationship between occupations and skills. When doing so we need to realize that the graph which we use as input is imperfect in terms of correctness and completeness. Looking at the false positives of the algorithm we can determine which skills are incorrect due to leak of correctness. Knowing which skills are not linked to occupations but should be can improve the current graph. In Table 4.3 a random sample of predicted skills which are not represented in the graph are shown. The node2vec algorithm could not only be used as a tool to quantify the relation between skills and occupations but also to improve the current graph by adding possible new skills.

An upon closer examination of the predictions of the Node2Vec model is displayed in Figure 4.3. In this figure the skills which are predicted for the ISCO code 2611 (Lawyers) are shown in the y-axes. Looking at the corresponding value on the x-axes the prediction probability is shown. Note that only predictions are shown where the probability of occurring is higher

| ISCO-Code | Occupation | Predicted Skill |
|---|---|---|
| 1341 | Child care services managers | children's physical development |
| 2261 | Dentists | prepare materials for dental procedures |
| 3251 | Dental assistants and therapists | dentistry science |
| 4110 | General office clerks | demonstrate professional attitude to clients |
| 5411 | Fire fighters | safety engineering |
| 6121 | Livestock and dairy producers | promote animal welfare |
| 7132 | Spray painters and varnishers | spray pesticides |
| 8344 | Lifting truck operators | hazardous materials transportation |
| 9111 | Domestic cleaners and helpers | provide lawn care |

**Table 4.3:** Predicted skills by node2vec that did not occur in the graph

than 0.5 (50%). Any value lower than 0.5 means that the probability that a given skill is not linked to the occupation is higher, thus the skill is not linked. Within the figure green bars denote that a predicted skill is correctly predicted, meaning that the predicted skill is in the knowledge graph. Blue bars depict skills that are predicted but do not occur in the graph. Notably, the unjustly predicted skills have a lower prediction probability compared to correctly predicted skills. After modeling the relatedness of skills and occupations in a quantitative manner one could use the results in two main ways. (1) ranking the importance of a skill for an occupation and (2) using the predictions to improve the current dataset. The first use-case helps in situations when job candidates need to be compared. If two candidates

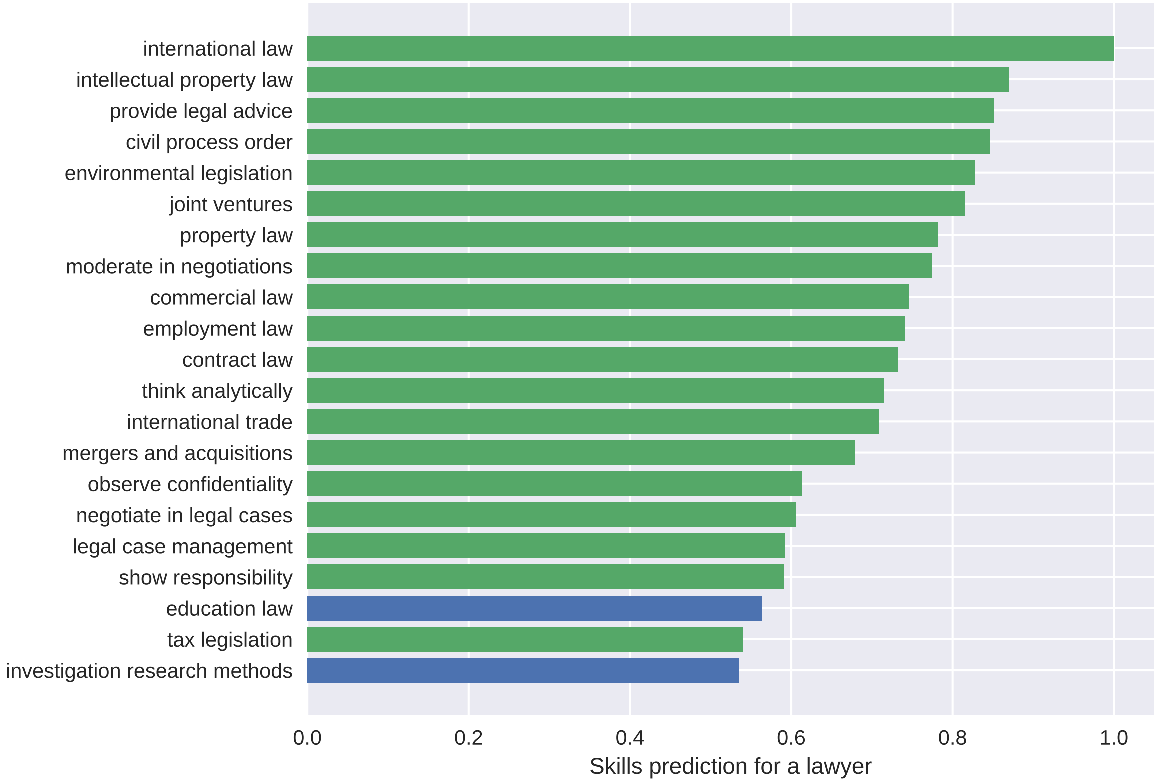


**Figure 4.3:** Predictions of the Node2Vec algorithm for ISCO group 2611 (Lawyers)

both have the same number of quantification requested for a job the skills that are offered could be ranked to determine which candidate has the most related skills for the job. The second use-case can ensure a higher coverage in the used knowledge graph. For example, looking at Table 4.3 one can see that dentists do not posses the skill "prepare materials for dental procedures" according to the current knowledge graph. This skill might be a good addition for the knowledge graph. By consulting domain experts skills can be efficiently added to enrich the current graph. In both cases the best result can be achieved by applying a machine driven approach with a human touch.

# Chapter 5

# Using Set-Based Node Similarity to Model Career Paths

Throughout a career multiple positions are fulfilled. As one develops their skills, new opportunities arise. Some people prefer to switch jobs when an opportunity presents itself due to chance, others do not. The second group likes to work towards something. When having a career goal in mind knowing how to reach it is key. There are a lot of factors in play when trying to reach this goal, one important one being the skills you posses. In this chapter we focus on how one could switch between occupations based on the skills.

According to the most recent data (2019) 1.1 million people switched occupation in the Netherlands [10]. When transitioning between jobs the difference between the old job and the new job cannot be too large. The gap between two jobs is too large if the skills required for the first job and the second job differ too much. Occupations that share a lot of skills are easier to transfer between since shared skills indicate shared tasks and responsibilities. To determine if one can transfer between occupations the distance between occupations needs to be quantified. This distance can be calculated using the jaccard distance, as elaborated upon in Section 2.2. Calculating the jaccard distance between two occupations is done by representing an occupation as the set of skills required for the occupation. This way two sets are compared with each other.

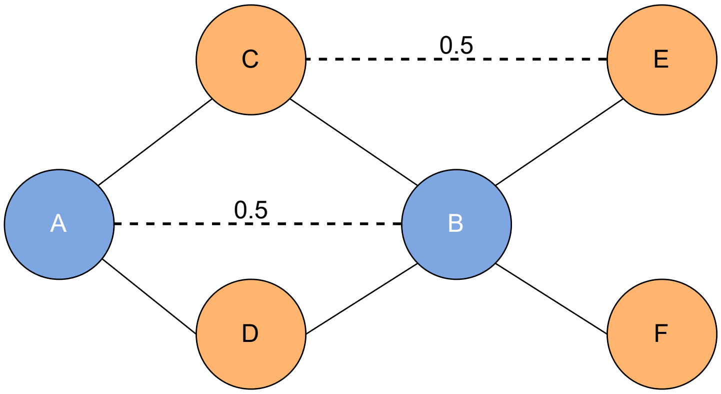


**Figure 5.1:** Jaccard distance in a graph where nodes {A, B} are occupations and nodes {C, D, E, F} are skills. Solid lines denote direct connections, dashed lines denote the jaccard distance.

The resulting score indicates the overlap in skills. Here, a high number represents a large distance, meaning that two occupations are not alike. A low distance will indicate that a similar skillset is required for both occupations, in this case the occupations are similar. In Figure 5.1 one can see an example graph where blue nodes represent an occupation and orange nodes indicate skills. Between the occupations $A$ and $B$ the jaccard distance can be calculated by counting the number of shared occupations ($C$ and $D$) and dividing the total number of skills connected ($C$, $D$, $E$ and $F$). This results in a jaccard distance of 0.5 between occupation $A$ and $B$. Calculating the jaccard distance can also be done between two skills, by doing so skills that are alike can be detected. An example of this is shown as well in Figure 5.1 between nodes $C$ and $E$. Note that the corresponding jaccard distance is 0.5 as well. In our dataset, as constructed in Chapter 3, a total of $120952$ links can be made between pairs of skills and pairs of occupations. From these combinations $89.3\%$ is between skills and $10.7\%$ of the connections can be made between occupations. To gain insight in the similarity of skills and occupations in the labor market one could look at the distribution of jaccard distances. This is shown in Figure 5.2.

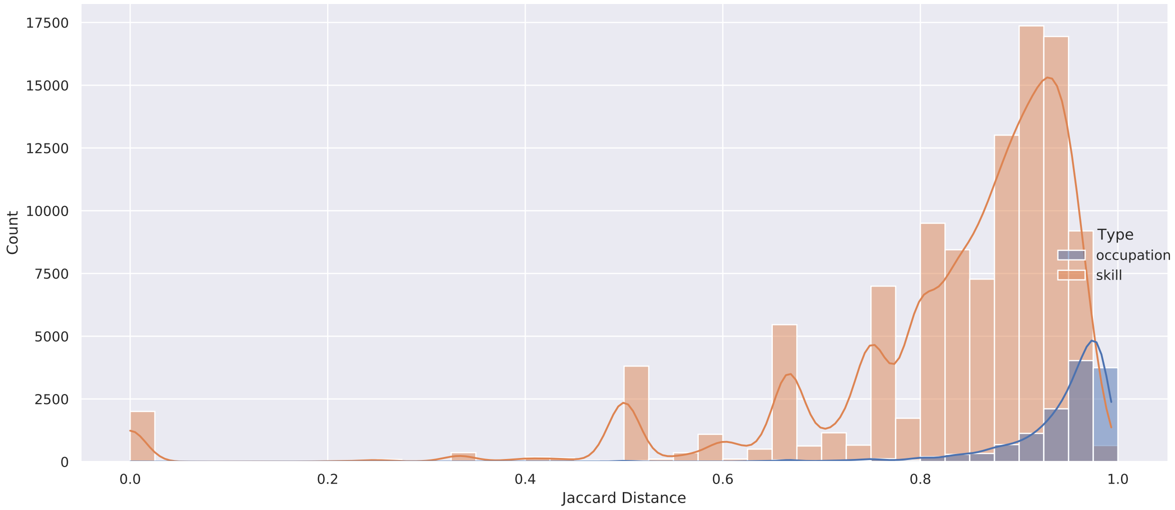


**Figure 5.2:** Distribution of the jaccard distance where the orange color represents the skills and the blue color represent the occupations

Looking at the distribution of jaccard distance one can see that skills are generally more similar than occupations. This becomes apparent when looking at the mean value of both distributions. For the occupations this value is around $0.96$ (x-axis) and for skills this value is around $0.88$. This height of the peak denotes the amount of observations. Having more observations does not mean that the observations are more similar, only that accuracy of the distribution is higher. More than 99% of the occupations have a jaccard distance between 0.8 and 1, this means that occupations require a distinct skillset. Both distributions are skewed to the left, meaning that the mean (average of the observations) is left of the mode (most observed value). In the distribution of skills a number of spikes are visible.

This can be explained by looking at the jaccard distance. When constructing the distance some fractions are more common than others. If half of the neighbors are shared the jaccard distance will be $\frac{1}{2}$. This result can be created in a number of different situations, for an example of this one can look at Figure 5.1. Other spikes occur at common fractures such as $\frac{2}{3}$ and $\frac{3}{4}$. In Table 5.1 a description of the distance distribution is shown. For both skills and

| | Skill | Occupation | Total |
|---|---|---|---|
| count | 107959 | 12993 | 120952 |
| mean | 0.825 | 0.938 | 0.837 |
| std | 0.163 | 0.070 | 0.160 |
| min | 0.000 | 0.000 | 0.000 |
| 25% | 0.800 | 0.928 | 0.800 |
| 50% | 0.875 | 0.960 | 0.888 |
| 75% | 0.923 | 0.977 | 0.933 |
| max | 0.985 | 0.993 | 0.993 |

**Table 5.1:** Statistics of the jaccard distribution

occupations the minimum distance is 0, meaning that a skill is shared by every occupation where the skills is connected to or that two occupations share every skill. An example is *Food service counter attendants* and *Hotel receptionists*, both share the same skillset and thus have a jaccard distance of 0. Skills with a distance of 0 are for example *Lop trees* and *Pruning techniques*. The highest distance found in the dataset is $0.993$, this corresponds with the occupations *Electronics engineers* and *Policy administration professionals*. They share at least 1 skill but are beside that skill the most different. The common skill is in this case *perform project management*.

Once the distance between every occupation is calculated the most efficient transition between every pair of occupations is evaluated. This is done by finding the most efficient path between the start node and the desired end note. Here the start node is the current position and the end note the desired occupation. When the distance between a pair of occupations is too high one might not be able to learn the required skills at once to start the new function. In Figure 5.3 an example of a transition can be found. Here we set a threshold for the maximum possible distance to cover to a value of $0.8$. If two occupations are further apart than $0.8$ the step is too large. In this example we start at node $W$ and desire to go to node $Z$. We are not able to directly transition between $W$ and $Z$ because the occupations are

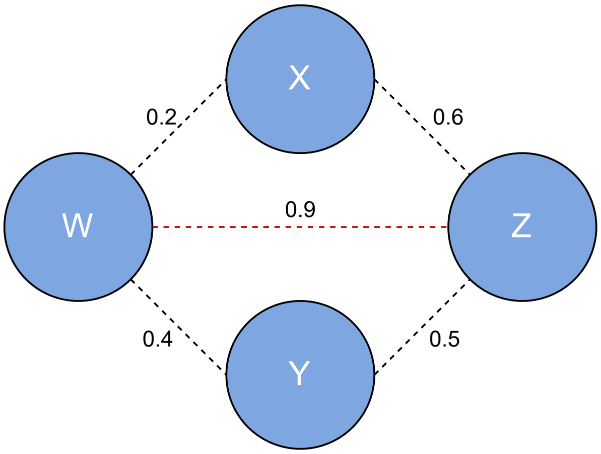


**Figure 5.3:** Distance between the occupations $\{W, X, Y, Z\}$. Black lines denote distances lower than $0.8$. Red lines denote distances higher than $0.8$.

not similar enough. Finding the most efficient transition can be done by finding the shortest path. Numerous shortest path algorithms are available for undirected graphs, for the purpose of this thesis Dijkstra's algorithm [12] is used. The shortest allowed path between $W$ and $Z$ in Figure 5.3 is by going over node $X$. This could be interpreted as when one holds position $W$ and wants to become $Z$ they have to become $X$ first.

A real world example is shown in Figure 5.4. Due to the COVID-19 pandemic a lot of people find themselves out of a job, especially for individuals that work in restaurants. Those people need to find a job as quickly as possible. Using the described model we can calculate which occupation has the closest proximity to cooks. By looking at the distance we find that "bakers, pastry-cooks and confectionery makers" are most similar.

Knowing how to efficiently transition between two occupations is an essential part of career pathing. This enables somebody with a specific career goal to reach it as frictionless as possible, looking from a skill-requirements perspective. Knowing were to go helps in continually advancing one's career. If skills are required for a next step but not yet possessed one can address this beforehand by reskilling when required.

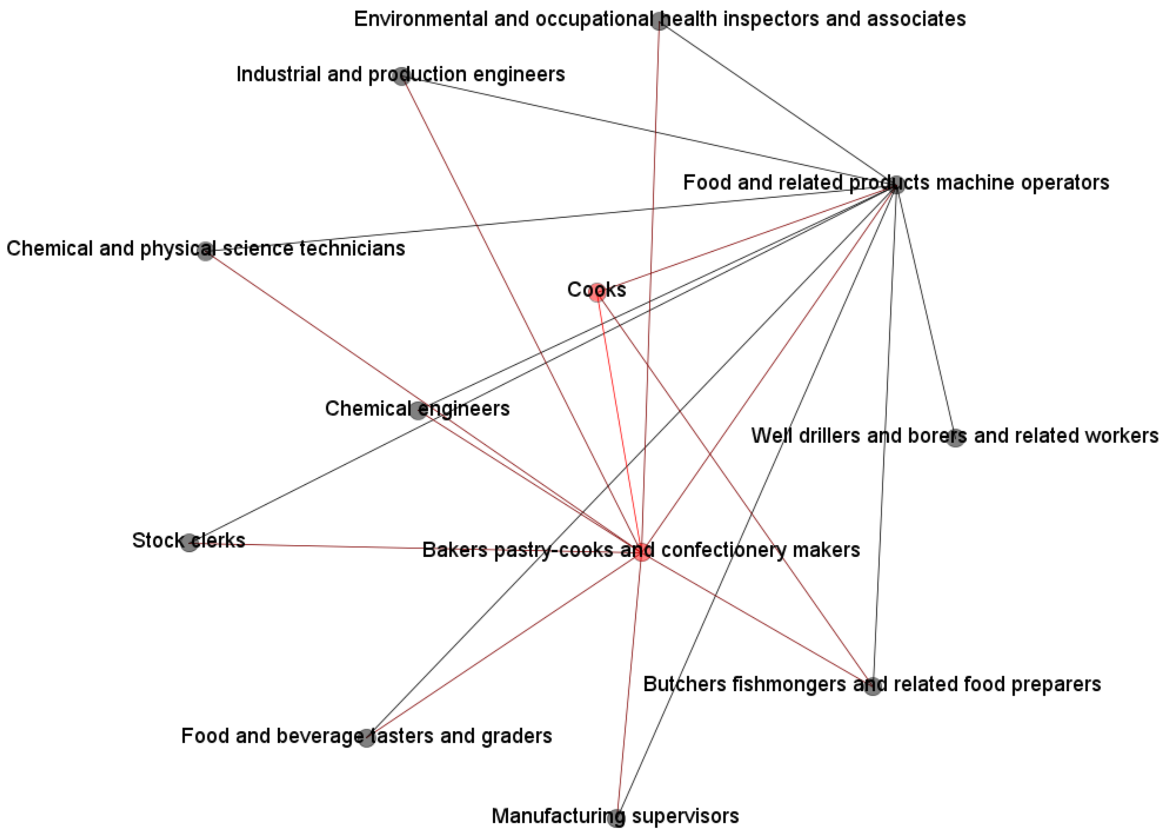

Environmental and occupational health inspectors and associates
Industrial and production engineers
Food and related products machine operators
Chemical and physical science technicians
Cooks
Chemical engineers
Well drillers and borers and related workers
Stock clerks
Bakers pastry-cooks and confectionery makers
Butchers fishmongers and related food preparers
Food and beverage tasters and graders
Manufacturing supervisors

**Figure 5.4:** The shortest path between the occupation cooks and the closest connected occupation, in this case Bakers, pastry-cooks and confectionery makers

# Chapter 6

# Using TF-IDF to Model Skill Relevance on Multiple ISCO levels

Individuals do not always know which job are going to do. Sometimes people only know that they want to work in a given industry. Acquiring relevant skills is difficult when the occupation that one is going to fulfill is unknown. In this chapter a method is proposed to determine what skills are the most relevant. This is done on multiple levels if the ISCO taxonomy.

Between different occupations skilled are shared. Some skills, such as *teamwork*, are required for a large number of occupations. Possessing this skill is often considered a basic requirement. Since this skill is required for so many functions it becomes almost trivial. Discovering relevant skills for an occupation can improve the match between a job candidate and the job posting. Whether a skill is relevant or not can be measured in different ways. For a skill to be relevant we define two criteria:

- A skill needs to be frequent within its context.
- A skill needs to be characteristic for its context.

Measuring relevancy can be done on the occupation level (ISCO level 4) or at an aggregation of occupations (ISCO level 1-3). The level of detail in the measurement is the context of the skill. Within the ISCO taxonomy 4 granularity levels can be used to determine relevance. Note that relevance does not equal importance in this thesis. Teamwork is important in a lot of contexts, but it might not be relevant in a specific context because it is required in almost all contexts. When comparing context alongside each other the skill is compared with all groups in the same ISCO level.

## 6.1 TF-IDF

To achieve relevance score the Term Frequency–Inverse Document Frequency (TF-IDF) [34] statistic is used. This statistic is chosen based in previous successful applications in [33], [38] and [11]. TF-IDF has proven to be a robust measurement that has become the standard for term-weighting schemes. A survey from 2015 showed that more than 80% of the recommender systems use this metric [4].

To understand how TF-IDF works the following terminology is required:

- term ($t$): the object that is of interest, in this case the skill.
- document ($d$): the context where the term is, here this is an ISCO group.
- corpus: the collection of documents.
- count of corpus ($N$): The number of document in the corpus.

TF-IDF consists of two parts, this first part is the Term Frequency (TF). This is the frequency of a word used in a given context. Here, the frequency is calculated as the count of a given terms divided by the total amount of terms within a document. If ISCO group 9999 has the a total of 10 skills and a skill occurs 3 times the TF of that skill is $0.3$. The term frequency is always between 0 and 1 due to this normalization. A higher frequency denotes a more relevant term. If a skill occurs in all ISCO groups with a high frequency this skill does characterize any given group. To counter that overall common skills become relevant for every context the Inverse Document Frequency (IDF) is used. This metric returns a high number if a term appears in a low number of documents. If a term is common the IDF will be a low number. The TF-IDF is constructed by multiplying the TF with the IDF, this is defined by the following formula:

$$TF - IDF(t,d) = tf_{t,d} \times \log\left(\frac{N}{df_t + 1}\right) \tag{6.1}$$

$tf_{t,d}$ = the frequency of $t$ in $d$
$df_t$ = the number of documents containing $t$
Note that denominator in the IDF is increased by 1, this is to ensures that the IDF never divides by zero.

For this TF-IDF based model skills and occupations from the knowledge graph are used. The count of the skills, which determine frequency, are based on the number of matches from the job posting. Here the matching process is followed as described in Section 3.3. The count of a skill is dependent on the occupation for which it corresponds in the graph. If a skill occurs 10 times in ISCO group 1011 and 5 times in ISCO group 1012 the count is aggregated when looking at the group 101. Note that in this example group 1011 and group 1012 are subgroups of 101. The TF-IDF score is calculated for every occupation by taking the term frequency from within a group, independent of other groups. The inverse document frequency is taken by comparing every group with every other group in the same level. Here group 1011 is compared with group 1012 but not with group 101 since this last group is not at the same level. The level of a group can be deducted when looking at the length of the group number.

The resulting score provides us with skills that are common in a given document but uncommon in all other documents. For the purpose of this document this defines relevance. The relevance can be calculated on every level of the ISCO taxonomy. In Table 6.1 the top 5 skills are displayed for the level 1 ISCO groups. In this table *Microsoft Office* is displayed for the groups *managers* and *Clerical support workers*, this skill seems to be fundamental in both groups. For this skill to score high in multiple contexts the frequencies need to be substantial in both bases. In the managers group the skill *Microsoft Office* has a TF of 9% and *Clerical support workers* has a TF of 5%. Those high frequencies counter for the IDF component of the metric and explain how this can be ranked highly in two categories.

Looking at a single ISCO level helps to understand what skills are relevant for that layer. To deepen our understanding we will look at the development of multiple layers of ISCO group 2. A visual representation of this can be found in Figure 6.1, here the top 3 most relevant skills are displayed. Due to space constraints only a subset of the ISCO groups is shown. From this figure one can notice the following observations:

1. Communication appears in a number of different forms.
2. Nursing professionals have the same most relevant skills as Nursing and midwifery professionals and almost the same most relevant skills as health professionals.
3. Specialized skills appear more at the bottom of the figure than at the top.

The terms *communication*, *communication sciences*, *communication studies*, *ICT communication protocols*, *manage online communications* and *communication disorders* seem to be closely related when evaluated by a human. Because these skills are defined as distinct skills each skill receives its own ranking this concept can appear multiple times. Looking at ISCO groups 2221, 222 and 22 one might notice that the top three skills are almost the same on these different levels. Within the health professionals group nursing professionals appear to be the most characteristic ones. The high frequency in nursing like professionals is probably the reason for this. If a high number of vacancies in group 22 are in group 2221 the skills which are relevant in that group will have a higher relevance in the parent groups. More specialized skills, such as dental studies, are more commonly observed in level 4 ISCO groups. This is due to the lower frequency of specialized skills when a lot of occupations are combined in higher level groups.

In Figure 6.2 a matrix is provided which states the relevance of difference skills in the major ISCO group "Craft and Related Trades Workers". In this matrix different occupations are shown in the y-axes. A subset of the ESCO skills are shown on the x-axes. For every of the skill, occupation pair the relevance score is calculated. If a skill has no link with an occupation the relevance is 0 since the term frequency is 0. In this figure the relevance scores are normalized between 0 and 1 by dividing the resulting TD-IDF score from the model with the maximum TD-IDF score. Looking at Figure 6.2 the intuition behind relevance is shown. If an occupation has only 1 skill linked, the relevance of the single skill is high. An example of this can be seen for the occupation Spray painters and varnishers only the skill industrial paint is relevant. An occupation such as food and beverage tasters and graders has multiple skills, the relevance per skill is lower.

After creating a model that is able to calculate the most relevant skills for any ISCO code we are able to make a distinction between skills. Doing so provides insights in which skills are generally considered to be specialized and which skills are common skills in a given industry. Besides a classification of the skills these insights can be used to monitor market trends. By calculating the most relevant skills at different moment in time the change in relevance-scores can be used as an indicator of future relevance.

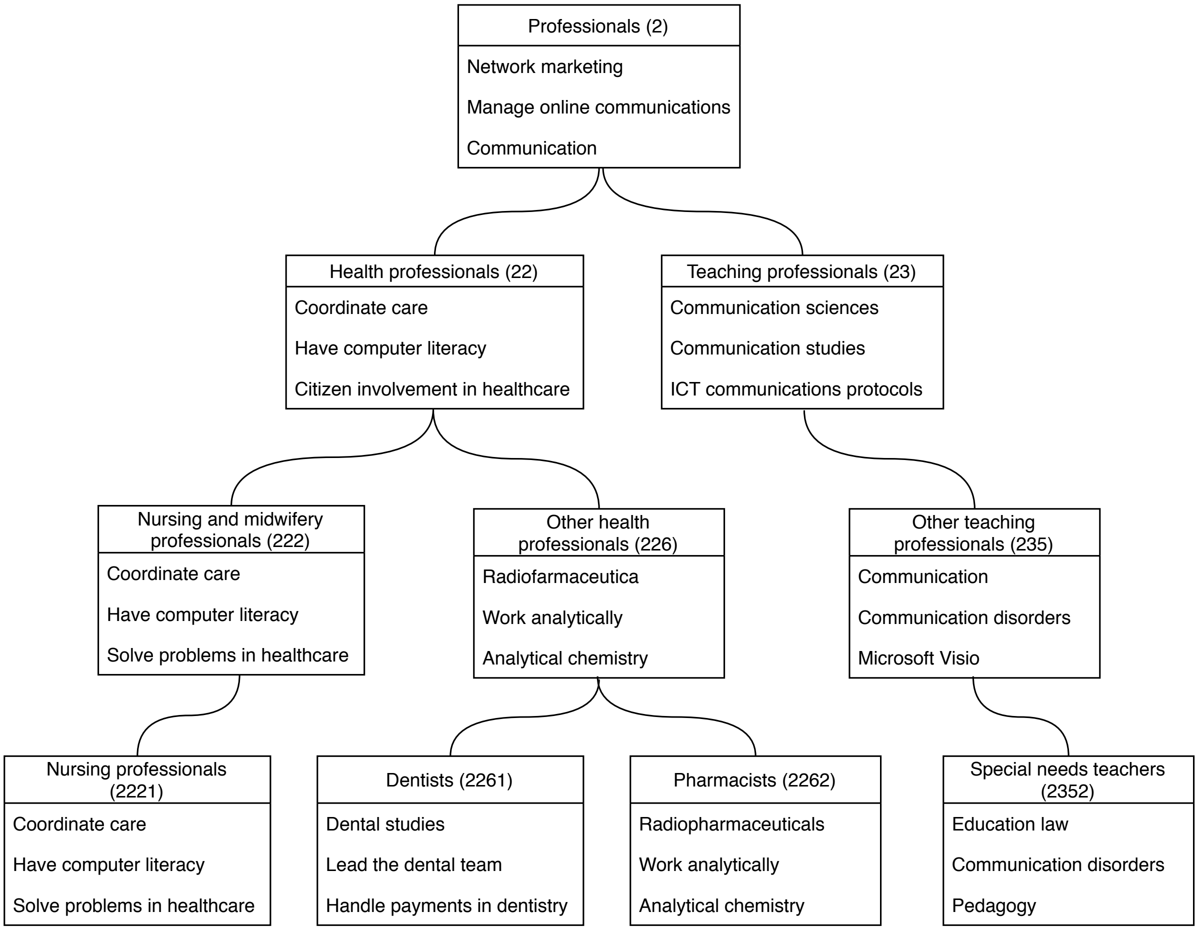


**Figure 6.1:** Three most relevant skills for multiple levels in major ISCO group 2

| | Managers | Professionals | Technicians and associate professionals |
|---|---|---|---|
| 1 | Microsoft Office | Network Marketing | Marker Making |
| 2 | Service-oriented Modelling | Manage Online Communications | Electronic Communication |
| 3 | Communication Principles | Communication | Service-oriented Modelling |
| 4 | Electronic Communication | Explain Accounting Records | Education Administration |
| 5 | Coordinate Patrols | Accounting | Manage Standard ERP System |
| | Clerical support workers | Service and sales workers | Skilled agricultural, forestry and fishery workers |
| 1 | Execute Administration | Security Panels | Leadership Principles |
| 2 | Perform Clerical Duties | Electronic Communication | Agricultural Information Systems and Databases |
| 3 | Microsoft Office | Create Solutions to Problems | Pruning Techniques |
| 4 | Education Administration | Execute Administration | Spray Pesticides |
| 5 | Human Resource Management | Recreation Activities | Lop Trees |
| | Craft and related trades workers | Plant and machine operators, and assemblers | Elementary occupations |
| 1 | Attend to Detail in Casting Processes | Mechatronics | Inventory Management Rules |
| 2 | Attention to Detail | Mechanical Engineering | Have Computer Literacy |
| 3 | Adobe Illustrator | Electrical Engineering | Carpentry |
| 4 | Adobe Photoshop | Operate Soldering Equipment | Place Concrete Forms |
| 5 | ML (computer programming) | Act Reliably | Operate on-board Computer Systems |

**Table 6.1:** Five most relevant skills per major ISCO group based in the TF-IDF matric

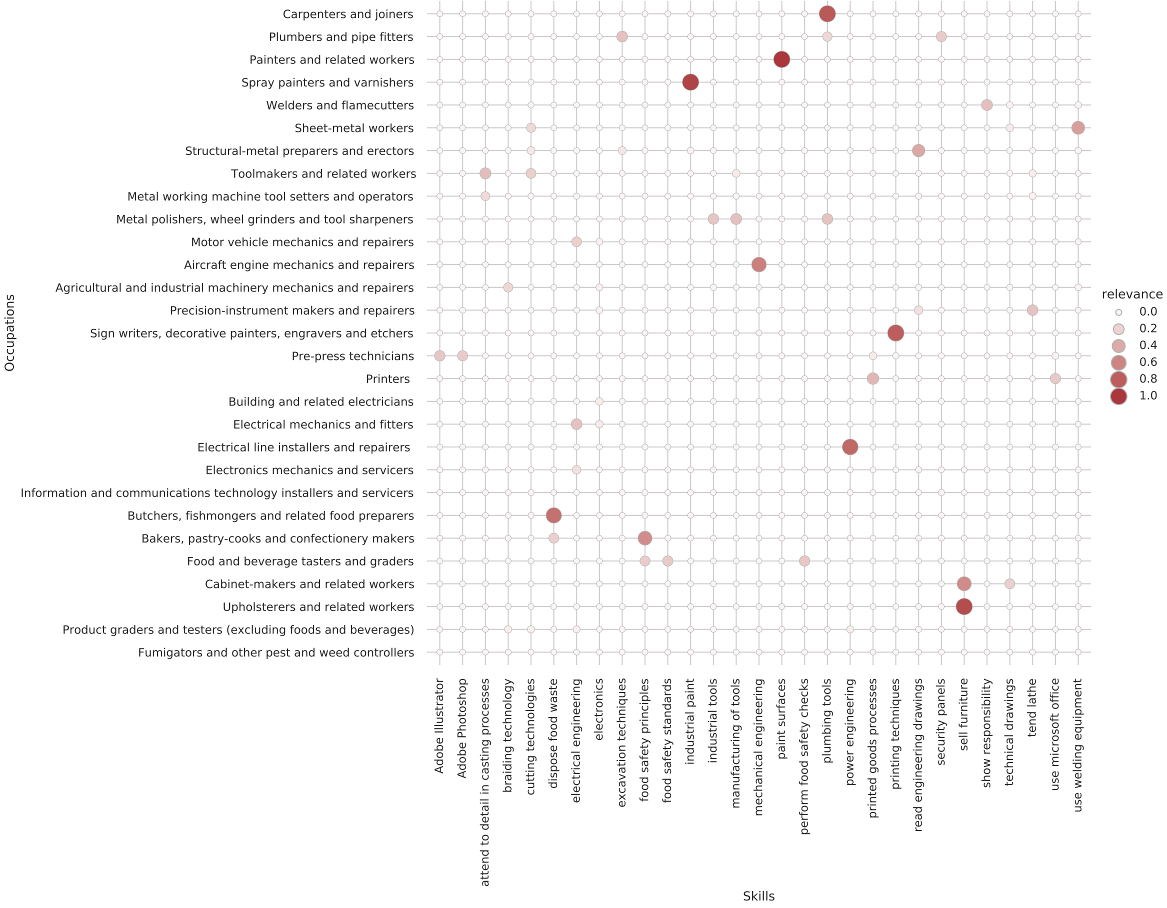


**Figure 6.2:** Relevance of skills for multiple occupations in the major ISCO group “Craft and Related Trades Workers”

# Chapter 7

# Conclusion

In recent years the labor market has changed drastically. This is mostly due to increased globalization, a growing working population and disappearing jobs due to digitalization. The COVID-19 pandemic has accelerated this change. This research aims to improve the fit between a job candidate and an occupation by focusing on skills. Modeling the relationship between occupations and skills can provide insights for new work seekers with an existing skill set. This can be used to find a job that matches the current capabilities of an individual. This problem is addressed in the first research question:

*How can we model the notion of relatedness between a skill and an occupation?*

Here three different link prediction methods are compared. For our dataset the best suited algorithm is called "Node2Vec", this algorithms is explained in Section 2.3.1. By using this method one is able to predict the relation between any skill and any occupation. Using this algorithm one can (1) find new skills for an existing occupation and (2) rank the known skills for an occupation.

When an individual is searching for a job knowing which occupations fit best is key. By modeling the relationship between occupations the most efficient path between jobs can be created. This helps if somebody already has a job but wants to transition to the next one. When switching jobs the skills required for the new job need to be known. If the skill gap between jobs is too large the switch might not be possible. To answer the second research question:

*Given two occupations, what is the most efficient transition based on skill similarity?*

In Chapter 5 a method is proposed to quantify the distance between jobs based on the required skills. Knowing the distance between a set of occupations gives the possibility to quantify reachable new jobs, here a reachable distance can be set. Meaning that if two jobs are too far apart the transition between jobs is not possible in one hop. Extending on this result, the most efficient route can be mapped between jobs by reaching a desired job with intermediate jobs in between. This career pathing strategy helps achieving a desired occupation with as little friction as possible.

For people that have to acquire new skills the most relevant skills will be provided for any job, industry or groups of industries. Research question three states:

*How can we determine what skills are the most relevant in any depth in the ISCO taxonomy?*

These skills have been calculated by taking the frequency of skills and the uniqueness of skills in a level of the ISCO taxonomy. Here, the uniqueness is high if a skill occurs more often in a group compared to other groups. The metric that reflects this intuition is called "TF-IDF". This calculation is done for multiple levels of detail in the job market. By doing so birds-eye view of the labor market is constructed where relevant skills can be detected. All these facets are used to answer the main research question of this research:

*How can the perfect fit between a job candidate and occupation be found based on skills?*

To create the perfect fit based on skills the skills gap between the current situation of the job candidate and the desired situation need to be as small as possible. To close this gap the current skills of an individual, current situation and desired situation need to be identified, doing so reveals a number of scenarios. If an individual does not posses any skills yet the most relevant skills for the desired situation need to be acquired. When the job candidate has a number of skills the relatedness between the skills and the desired occupation can be modeled. By doing so the occupation which is most related to the skillset can be found. If the individual is already employed but seeks a change in employment the most efficient transition between the current occupation and desired occupation can be calculated.

# Chapter 8

# Discussion & Future research

In this work a framework is proposed which is ontology agnostic, as long as the job postings are labeled with the same labels as the used selected ontology this method can be used successfully. This thesis uses the a combination of the ISCO and the ESCO frameworks, doing so has a number of advantages. These frameworks are available in a large number of languages, reorganized standards and freely available.

Other frameworks could be used as well, where ESCO is widely used in Europe the O*NET framework [30] is often referred to as the de facto in United States. Substituting the ESCO for the O*NET skill framework (or any other framework) is possible for the proposed method. When multiple languages are used to enlarge the set of available job postings, skills are linked by the identifier of the skill. Doing so can introduce a bias based on the input language. Different languages correlate with different cultures, in a language independent framework this bias needs to be taken into account. By using vacancies of different languages the underlying cultural differences can cause unforeseen, and possible unwanted side effects. The fine grained information which is constructed by looking at the vacancy level could be disrupted due to these differences.

Verifying results of any experiment requires a reference for the truth. Using this work the minimal path between any two occupations can be calculated. These calculation are based on a distance measure used. To know if the presented path is indeed the shortest one additional data is required. This verification data can be used to test the quality of the predictions made, without this external truth the accuracy of the predictions remain unknown. Unfortunately, no such data is available at the time of writing. The same problem is present for the prediction made regarding the relevance of skills.

The outcome of any research is heavily dependent on the available data. In the case of this research this data is preprocessed in a number of steps, one of which is the skill matching step. In this step the list of candidate skills is compared to the predefined skills. This process in its current form is sub-optimal and could undergo further improvement. The two main improvements are (1) lowering the number of false positive matches and (2) enforce a 1-to-1 match. By only looking at the structure of a word, two skills could be matched while the meaning of the skills are different, an example of this is meteorology and metrology. Here both skills are only two letters apart, which results in a match using the current matching technique. Since the meaning of these skills are different this match can be classified as an false positive. The second problem is that one candidate skill can match multiple predefined skills. A candidate skill like "communication" will be matched, among other things, to "communication studies" and "communication sciences". As a result of this a co-occurrence is created between a number of skills that match to the same candidate skill. This disrupts the outcome of the conducted experiments in this research. Examples of this can be found throughout every sub question. In the link prediction task the co-occurrence of a skill is predicted if this skill is not part of the primary skills for the occupation. In the node similarity the number of overlapping skills will be enlarged. This is because instead of one matching skill two skills will match if both skills are linked. For the skill relevance the co-occurring skills will gain a similar importance, this effect can be seen in Figure 6.1.

The output of any model is only as good as the data which is used. In the case of this research the initial dataset is created by taking $600.000$ job postings, extracting the candidate skills and matching these skills with skills from an existing framework. Extracting correct skills from job posting has a number of challenges. The job postings consist of unstructured text, detecting the correct candidate skills is a non-trivial task. Often tasks are described in text, this makes it more difficult to correctly match candidate skills to verified skills. This problem could be averted by using resumes rather than job postings. The text in resumes is often better for extraction tasks. If this research was to be developed further the skill matching procedure needs to be improved. Most of the data related shortcomings can be traced back to this matching process. An improvement over the current process could be achieved by using word embedding.

Having a flawed knowledge graph as a result of sub-optimal prepossessing does not invalidate the methods used. Whichever approach is used to create a knowledge graph, the outcome will never be perfect [6]. It will always hold that all models are wrong, but some are useful [9].

# Bibliography

[1] Charu C Aggarwal. *Data mining: the textbook*. Springer, 2015.

[2] İ. Semih Akçomak, Lex Borghans, and Bas ter Weel. Measuring and interpreting trends in the division of labour in the netherlands. *De Economist*, 159(4):435–482, Dec 2011. `https://doi.org/10.1007/s10645-011-9168-3`.

[3] Pol Antràs, Luis Garicano, and Esteban Rossi-Hansberg. Offshoring in a knowledge economy. Working Paper 11094, National Bureau of Economic Research, January 2005. `http://www.nber.org/papers/w11094`.

[4] Joeran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. Paper recommender systems: a literature survey. *International Journal on Digital Libraries*, 17(4):305–338, 2016.

[5] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data: The story so far. In *Semantic services, interoperability and web applications: emerging concepts*, pages 205–227. IGI Global, 2011.

[6] Antoine Bordes and Evgeniy Gabrilovich. Constructing and mining web-scale knowledge graphs: Kdd 2014 tutorial. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, page 1967, New York, NY, USA, 2014. Association for Computing Machinery. `https://doi.org/10.1145/2623330.2630803`.

[7] Lex Borghans, Bas Ter Weel, and Bruce A. Weinberg. People skills and the labor-market outcomes of underrepresented groups. *ILR Review*, 67(2):287–334, 2014. `https://doi.org/10.1177/001979391406700202`.

[8] Nicole Bosch and Bas Weel. Labour-market outcomes of older workers in the netherlands: Measuring job prospects using the occupational age structure. *De Economist*, 161, 06 2013.

[9] George E. P. Box. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976.

[10] centraal bureau voor de statistiek. De arbeidsmarkt in cijfers. `https://www.cbs.nl/-/media/_pdf/2020/18/dearbeidsmarktincijfers2019.pdf`.

[11] Ondrej Chum, James Philbin, Andrew Zisserman, et al. Near duplicate image detection: Min-hash and tf-idf weighting. In *Bmvc*, volume 810, pages 812–815, 2008.

[12] Edsger W Dijkstra et al. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271, 1959.

[13] european commission. Esco handbook. https://ec.europa.eu/esco/portal/document/en/0a89839c-098d-4e34-846c-54cbd5684d24.

[14] Eurostat. Labour market transitions – annual data. https://ec.europa.eu/eurostat/web/lfs/data/database.

[15] World Economic Forum. The future of jobs report 2020. World Economic Forum, Geneva, Switzerland, 2020.

[16] Burning Glass. Lens. https://www.burning-glass.com/products/lens-suite/.

[17] Burning Glass. Vacancy data. https://www.jobdigger.nl/.

[18] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.

[19] international labour office. International standard classification of occupations. https://www.ilo.org/public/english/bureau/stat/isco/.

[20] international labour office. International standard classification of occupations - structure, group definitions and correspondence tables. https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/---publ/documents/publication/wcms_172572.pdf.

[21] Paul Jaccard. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50, 1912.

[22] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[23] P Richard G Layard, Richard Layard, Stephen J Nickell, Stephen Nickell, and Richard Jackman. *Unemployment: macroeconomic performance and the labour market*. Oxford University Press on Demand, 2005.

[24] Chin Yang Lee. An algorithm for path connections and its applications. *IRE transactions on electronic computers*, (3):346–365, 1961.

[25] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.

[26] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[27] Nils J Nilsson. Stuart russell and peter norvig, artificial intelligence: A modern approach. *Artificial intelligence*, 82(1-2):369–380, 1996.

[28] OECD. Employment by job tenure intervals - average tenure. https://stats.oecd.org/Index.aspx?DataSetCode=TENURE_AVE.

[29] OECD. Ftpt employment based on national definitions. https://stats.oecd.org/Index.aspx?DataSetCode=FTPTN_D.

[30] O*NET. O*net online. https://www.onetonline.org/.

[31] Heiko Paulheim. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web*, 8(3):489–508, 2017.

[32] Gang Peng. Do computer skills affect worker employment? an empirical study from cps surveys. *Computers in Human Behavior*, 74:26 – 34, 2017. http://www.sciencedirect.com/science/article/pii/S0747563217302510.

[33] Juan Ramos et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer, 2003.

[34] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

[35] Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

[36] Robert Tarjan. Depth-first search and linear graph algorithms. *SIAM journal on computing*, 1(2):146–160, 1972.

[37] Textkernel. Extract. https://www.textkernel.com/nl/solution/extract/.

[38] Bruno Trstenjak, Sasa Mikac, and Dzenana Donko. Knn with tf-idf based framework for text categorization. *Procedia Engineering*, 69:1356–1364, 2014.

[39] Douglas Brent West et al. *Introduction to graph theory*, volume 2. Prentice hall Upper Saddle River, NJ, 1996.

[40] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, Sören Auer, and Pascal Hitzler. Quality assessment methodologies for linked open data. *Submitted to Semantic Web Journal*, 1:1–5, 2013.