DOCTORAL SCHOOL OF BUSINESS INFORMATICS

THESIS SUMMARY

Neusch Gábor Loránd

Labor Market Data Warehouse Design for Competence Trends Analysis

Ph.D. dissertation

Supervisor:

Borbásné Dr. Szabó Ildikó

associate professor

BUDAPEST, 2021

Institute of Information Technology

THESIS SUMMARY

Neusch Gábor Loránd

Labor Market Data Warehouse Design for Competence Trends Analysis

Ph.D. dissertation

Supervisor:

Borbásné Dr. Szabó Ildikó

associate professor

Table of Contents

1. RESEARCH GOALS AND RESEARCH QUESTIONS 1

1.1. RESEARCH QUESTIONS 4

1.2. RESEARCH FRAMEWORK 7

1.3. THE SIGNIFICANCE AND POSSIBILITIES OF THE RESEARCH 8

2. METHODOLOGY AND RESULTS 10

2.1. METHODOLOGY OF RESEARCH RELATED TO THE FIRST RESEARCH QUESTION 10

2.2. RESULTS OF THE FIRST RESEARCH QUESTION 11

2.3. METHODOLOGY OF RESEARCH RELATED TO THE SECOND RESEARCH QUESTION 15

2.4. RESULTS OF THE SECOND RESEARCH QUESTION 16

2.5. METHODOLOGY OF RESEARCH RELATED TO THE THIRD RESEARCH QUESTION 17

2.6. RESULTS OF THE THIRD RESEARCH QUESTION 18

3. REFERENCES 20

4. PUBLICATIONS 24

1. Research goals and research questions

According to Grundke and co-authors significant changes are going to take place in the labor market in the future as digital technologies continue to expand resulting in the complete automation of certain jobs while in case of others the tasks and the nature of the job will go through immense transformation. (Grundke *et al.*, 2018). Those working in the service sector, whose tasks require cognitive skills and knowledge primarily that are easy to turn into algorithms, may lose their jobs but at least they have to take into consideration the effects of comprehensive changes generated by automation. It is very likely, whatever course the future may take, that the 21st century man has to reinvent himself in every several years again and again; he has to acquire new skills and knowledge perpetually in the midst of escalating technological development as "new jobs arise due to the digital revolution" and "the content and nature of individual jobs as well as their scope of required skills keeps changing." (Grundke *et al.*, 2018, p. 6).

The labor market's demand is highly volatile it is changing at an extreme pace. Competence and knowledge become obsolete quickly and newer and newer intelligence is required. Automation gains more and more ground in different areas; human workforce is replaced by robotization and mechanization connected to the application of machine learning and artificial intelligence in terms of tasks that require cognitive skills. Against the labor market's rapid demand change there are institutions on the supply side that cannot easily adapt but reactively try to "measure up" to these demand changes. Most training institutions is characterized by rigid, hierarchic structure, bureaucracy, as well as exposure to internal and external political environment. In general syllabi are put together exclusively based on educators' own – necessarily limited – experience and biases, etc. Due to the lack of objective information, even with the best of intentions, curricula and syllabi cannot reflect market needs[1] correctly. Training plans go through the accreditation process that also takes place in huge, bureaucratic institutions. This, again, makes quick response to demand change difficult.

[1] Certainly, in many cases, this is not the (only) purpose; for example, in case of foundational subjects or generally in undergraduate training when the goal of education is not to strengthen profession specific knowledge, but to reinforce general competencies that serve as a basis for specific knowledge. Undoubtedly, beside these, many other perspectives are to be taken into consideration during curriculum development such as pedagogical, social, political, etc.

Beyond the above, in case of higher education institutions a further difficulty is posed by the fact that it is not the current market needs that need to be taken into consideration, but actually future needs should be considered, since the student whose training plan is put together in the given year is only going to enter the labor market years later. However, preparation for future needs is burdened by immense uncertainty precisely because of the extremely rapid change of labor market needs.

So, in the midst of extremely rapid changes in an immensely volatile environment when technology develops at an exponential rate – even though the pace forecasted by Moore's law seems to slow down by now (Waldrop, 2016) – it is extremely difficult to make a long term forecast for the next 10-20 years and give advice to institutes of higher education based on which they can develop curricula that adapt to the labor market demands. Those who still attempt to give such advice highlight that student competences to be improved primarily are the ones connected to emotional intelligence and organizational skills (Beck and Libert, 2017) and to provide students with such a strong foundation on which they can easily build when they learn new skills and knowledge that line up with labor market demands.

Beyond establishing the most important basic skills that students can use in the long run, in order to ensure short term competitiveness educational institutions should keep curricula adapted to labor market demands on an ongoing basis that are influenced by several factors and their forecast is by no means trivial. The greatest difficulty of those concerned in this matter, directors of studies, is that the change of demands is invisible in advance. If they were able to obtain an objective picture of how demand for different competences will develop in the labor market in time, they would be able to make conclusions in terms of future trends. The most important piece of information required for such a forecast is the momentary representation of labor market demands recorded in time series. This can be understood as a sum of "snapshots" of the sets of competencies in demand in the given moments of time.

In my dissertation I intend to focus on the coordination of the competence demand with the higher education supply in the short run by outlining a framework that can be used for the real time analysis of labor market trends and data describing demand. I attempt to develop the concept of a labor market data warehouse where information about labor market demands that appear in job listings can be accumulated. I examine different methods of extracting information reflecting competence demand that appear in job

listings either implicitly or explicitly. Based on such an information source – perhaps complemented by data from other sources, market and industry analyses, reports, such as *Gartner's Hype Cycle*, etc. – executives of higher education could continuously adapt course syllabi to labor market demand, and they would be able to offer opportunities for students to acquire competencies that they can successfully capitalize on in the first years of their career. In the midst of the current technological development causing perpetual discontinuity when *artificial intelligence* and *machine learning's* disruptive effect keeps transforming the labor market, such information could mean a substantial competitive advantage for an institute of education or, on a higher scale, even for a national economy.

So, my broader research refers to the problem of what methods can be used and how can one analyze and forecast labor market demand and examine how demand and supply correspond. Furthermore, to coordinate them, to reach that training completion requirements would reflect more and more the competence requirements that appear in job listings. As described above, by supply I do not mean the sum of individuals (employees), but some sort of output quality provided by a training while on the demand side I speak of competence (as resource) need. Since the development of the complete research would go beyond the limits of a Ph.D. dissertation therefore, I present the applicability and feasibility of the framework that serves as the foundation of the research. That is, the dissertation and its research questions pose to find solutions for only a portion of the complete problem area; focusing on the demand side of the labor market, namely, how can one effectively explore and retain information signaling competence demand, thereby providing basis for analyses that later make the better alignment of supply possible.

The presented framework – following its implementation – may help higher education executives – primarily teachers and directors of studies – in creating such syllabi through which competencies offered to students are still going to be valid and marketable when those students who start their university career today enter the labor market. This primarily requires syllabi to be made up of elements that ensure that competencies defined based on up-to-date labor market demand are passed on to students.

During data collection I used the basic assumption that demand for competencies available within the framework of my research is best reflected in job listings. Since the human resource need, that is, the kind of workforce a particular company needs is best described and communicated to potential candidates based on the task to be done or the

position that needs to be filled therefore, the companies' competence needs necessarily have to appear in job listings.

This assumption is reinforced in literature by Pitukhin and co-authors for instance who write that most professional requirements that the employer sets for a candidate appear in job listings. (2016, p. 2028). Wowczko also highlights that along with the emergence of online recruiting a huge amount of potentially useful information is available for researchers in reference to requested competencies. (Wowczko, 2015, p. 34). Zhao and co-authors (2015, p. 4013), researchers of CareerBuilder.com have found upon examining a sample taken from half billion job listings is English that ninety percent of them contain expressions that can be accepted as competencies from this dissertation's point of view. As opposed to the previous Nasir and co-authors (2020) mention that listings they processed rather emphasize vacancy information and they detail required skills less.

1.1. Research questions

In my dissertation I intended to outline the blueprint of a system and to support the feasibility of its most important modules that serve as the foundation of the whole solution whose main purpose was to provide such information that shows a picture of jobs, required competencies, certifications and trainings that appear in job listings and their evolution in time in the form of reports, analyses, etc. Obviously, the implementation of such a system is not a one-man task. The framework of the dissertation provided an opportunity to give a theoretically established suggestion for data storage, start the data collection, and based on that examine methods that can be used to identify competence elements as well as related jobs in job postings. I have defined my research questions accordingly.

The system's primary input data are job listings posted to online job search portals that require storage because of several problems. On the one hand, since they come from an internet source their availability cannot be guaranteed for the whole cycle of the analysis. For instance, in case of a job listing when it is no longer relevant for the advertiser – that is, the position has been filled – it is taken off the portal, but the information it contains is still important for us, because we also need data from the past for the planned analyses. On the other hand, since job listings are noisy, they need to be prepared for analysis. Preparation is necessary because information that serves as the basis for our analyses are

not always present in job listings in a structured form or even explicitly, so, they have to be extracted from the text and after the pre-processing has taken place the extracted information is worthy for storage. So, my first research question examines how the data storage solution should be developed and what kind of data should be collected.

1. Research question: What is the best tool (data storage platform) to store data collected from online sources – extremely heterogenous, large amounts of unstructured data – job postings and related economical and statistical facts in a way that based on information recorded this manner those involved can be provided with valuable analyses that show the evolution of competencies in demand on the labor market chronologically and along other dimensions.

1.1. What data fields should be collected and what can be the source of data? In the dissertation I examine technologies that can be used for data collection and present the implementation of the "web scraper" that was used to collect data for the experiments done to create the basis for the system.

1.2. What points need to be considered when choosing a data storage platform? In the theoretical part of the dissertation, I present the considerations that can serve as the basis for choosing the best system for the present problem. I have examined which data storage structure serves best the purposes of the system intended to be built. That is, which storage technology serves the purpose better, is it a traditional data warehouse solution or a data lake popular in big data environment. Furthermore, based on the points outlined I compare specific data storage architectures in the theoretical part of the dissertation with the purpose of supporting the correctness of the outlined conception by comparing the points that have served as the basis for my choice.

1.3 In the dissertation I also explain how the collected data should be loaded and organized into schemas. Is it at all necessary to define schemas for data storage as in the case of widespread data warehouse architectures, where usually through an ETL (extract, transform, load) process the cleared and organized data are loaded into a predefined relational data model? If yes, how should the schema look like? Or, is it more efficient – as in the case of a data lake – to store data in the form as they were collected from the source – for instance, in a distributed file system – while logic is applied during analysis and processing that can take place with the help of a *MapReduce* algorithm or a *Spark* job?

In summary, it is an important question to examine which approach serves the purpose best, a data warehouse, a data lake, or perhaps a hybrid solution? A further related question is whether we should store the job listings' data in predefined schemas and build an *ETL* process to clear and load data or is it better to store data in their collected form and apply logic during analysis if it is necessary.

After finding the data needed and the best storage solution for them, I have done the processing of the collected data in order to establish the further sections of the research where the goal is to build the connection between the labor market demand and education supply and to create a foundation for their harmonization later. The assumption of the thesis is that connection can be built based on competence therefore, one of the goals of this dissertation is to identify competence elements on the demand side that appear in job postings.

2. Research question: What kind of information technology solutions can automatically identify competence elements that appear explicitly in free text descriptions of job listings in the case of this present dissertation, but similarly also in process models' task descriptions? What level of accuracy can be reached with the different solutions?

Job listings or process models' task descriptions usually contain that most important list of competences that the employer needs to fill the position described in the job posting or to complete a given task. However, the identification of the n-grams representing these competences are not trivial in the free text corpus. During the exploration of the research question, I am looking for a solution as to how, by what means is it possible to identify these n-grams – that appear explicitly, and represent competence or knowledge – in the corpus of job listings.

Competence identification can take place with text mining tools. This process can be supported by using outside sources and these sources can aid the identification of the elements in job listings as a competence-dictionary. Such outside sources can be for instance the skill pillar of the ESCO ontology or the STUDIO ontology that I describe in detail in the dissertation.

In the dissertation I meticulously examine how job listings' underlying, hidden, implicitly present competence elements can be explored. This issue what my third research question is about.

3. Research question: What methods or technologies can be utilized to explore implicit competence elements that do not appear directly in job listings, but are relevant in the context of the posting? What data sources should be leaned on to identify these hidden objects?

Job listings highlight such competencies that are needed to fill the given position and the same is true in case of process models, in terms of a job. These expressions that appear explicitly – semantically (based on the connections between them or based on their meaning) or simply on a statistical basis (e.g.: joint occurrence, the distance of each element) – can also refer to such, explicitly not present, latent competencies that could be equally relevant in the context of the given position or job.

1.2. Research framework

From the point of view of the goal of the thesis the most important data are competence requirements, so, one of the most important milestones of the research include the identification, collection and semantic enrichment of these competencies from the input data, in a position level granularity (that is given by the job postings). Accordingly, a major part of my dissertation discusses the possibilities of the identification of competences that appear explicitly in job listings as well as competencies related to them implicitly.

Labor market supply is made up of numerous different groups like beginner employees – among them fresh graduates, experienced employees who are changing workplaces, or those who are re-integrating, re-joining the system for various reasons, etc. From this perspective the research focuses on the higher education sector, more specifically fresh graduates in accordance with the final purpose of the dissertation that is to provide educators with such information based on which output requirements can be filled with content that fresh graduates can sell on the market.

When creating the framework, in order to simplify the problem, I narrowed down the circle of examined job listings to the IT sector and I also filtered the content of outside ontologies accordingly – that I use as a mock dictionary to identify competencies. Certainly, if the outlined proof of concept proves to be feasible, then the solution is easily adaptable to other fields as well.

In the dissertation only those areas of the full research field are elaborated that relate to the demand side of the labor market, more specifically the concept of the labor market "data warehouse" and how related competences can be extracted from job listings and how different semantic sources can be involved in processing in order to extract implicit information.

Because I examine the correctness of the outlined concept and I chose to use English job postings that were targeted to the UK labor market. I opted for English because it is "the language of IT" and on the other hand, because the name of numerous technologies and technical terms are not widely used in Hungarian translation. Furthermore, the processing of the Hungarian language (due to its peculiarities) is more complicated and requires more resources than English so, the development of a *PoC* is more reasonable to use the latter. Finally, there are significantly more relevant job listings available on the Internet in English this way, the cardinality of input data can be increased compared to the case of Hungarian processing. For the above reasons English sources are more suitable for the examination of the correctness of the concept while if the feasibility of the outlined model is proven, and there emerges a need for its results, then in the later stages of the research the system can be equipped to manage the Hungarian language. It is another assumption of the dissertation that the need for competences for cutting edge technologies emerges in English speaking markets first; I explain this assumption further in the subchapter of the dissertation dealing with the possibilities of the future research.

1.3. The significance and possibilities of the research

In the past years numerous researches has started about competences in several disciplines. Psychology, neurology, and economics are equally examining the subject from different angles. From an economics perspective the development of new competencies is an investment in human resources and as such we certainly wish that it would at least return but more so, that it would produce profit. From the employers' point of view this profit manifests in the competitiveness of the company while from the educational-political perspective it becomes visible through the given country's economic performance. It follows from the previous that we would like to control our resources we invest in competencies and this control can be achieved best by strict and consistent planning.

It is especially important when educational politicians make decisions to take into consideration what knowledge and skills are going to be required in the short and medium term in the labor market and how to bridge the increasing gap between “the skills needed in a radically transforming economy and work supply” (Fazekas, 2017, p. 6). Therefore, the knowledge of how competence needs change in time can bear priceless information value for decision makers involved in politics or education. Accordingly, depending on the competencies’ abstraction level different levels of decision making can be supported using the framework put forward in the dissertation. The solution outlined in the thesis is aimed especially at the support of directors of studies, but on a higher level the results can be useful for political executives as well.

The infrastructure, however, that I wish to outline, is not available in a widely accessible way that justifies the relevance of the dissertation and the research. In her research Wowczko (2015) found that even though the consideration of the coveted skills in syllabus design is indispensable in terms of the harmonization of labor market demand and supply earlier researches tended to exclusively focus on the degree requirements of jobs that is of course important, but it is not sufficient. This present dissertation is one step of the research that was launched to fill this void.

Even though directors of study are primarily targeted by the solution outlined in the dissertation, the information provided by the framework could benefit students directly. For instance, different learning paths can be suggested to them. This is possible in different ways. On the one hand, when a student knows what kind of work they want to do in the future, that is, they choose a certain career and they orient toward that then such a unique learning path can be outlined for them that, by following it they can precisely and purposefully improve the skills and acquire the knowledge that suits best the coveted career. On the other hand, if we can detect how a student’s knowledge, studying in a particular major, is lacking then we can provide them with a list of jobs they could effectively fill after finishing the given major complemented by a list of particular competencies to be improved and specific knowledge gaps to be filled in order to be able to successfully apply to a certain position. Knowing the evolution of competence demand could be equally important for employees, since under the banner of lifelong learning they also have to keep their set of competences up to date against extremely quickly changing requirements, in order to remain competitive in the market. Since the development of a new competence also creates significant opportunity costs besides

specific material expenses therefore, it should not be incidental what individuals invest their resources into.

In summary, it can be said, that good or bad investment decisions in this area can impact every level of the economy in terms of individual, corporate, national as well as regional competitiveness and performance. The framework outlined in my dissertation can provide invaluable information to foster good decisions.

2. Methodology and results

2.1. Methodology of research related to the first research question

In case of the first research question the goal was to choose the right decision criteria for the data storage architecture and proposing the best architecture model based on those points. During the work related to the dissertation I started automatized data collection, I developed a possible data model and by examining the selection points in detail I made a recommendation for the implementation of the suggested data storage architecture. From this point of view my research is exploratory. “Exploratory research is done for three reasons: to ensure a better understanding of the subject, to test the feasibility of a later more thorough research, to develop methods that further research can apply” (Varga, 2014, p. 4; Szabó, 2000).

Within the bounds of a smaller software development project I examined tools available for data collection, and I have chosen a web scraper system called *Scrapy* that is an open source application written in *Python* and provides all the infrastructure needed for the task. The user simply has to describe the “business logic” – what different elements of the collected documents mean and in what form the user wishes to store them – using *CSS Selectors* or *XPATH* in the so-called *spider* classes.

Also, during the examination of the first research question I defined the scope of the data to be stored that are indispensable in reaching the final goals and I outlined a possible relational schema that can be implemented if a relational data storage solution is chosen.

I developed the architecture selection criteria using secondary research and literature review. The most important examined criteria included classic attributes such as speed, availability, scalability, solution costs, feasibility and available support. Beyond these, I

have compared the solutions based on requirements resulting from the idiosyncrasies of data structure and *CAP theorem* (*consistency*, *availability*, *partition tolerance*) consequences. I have also examined the possibilities of the applicability of time series databases.

2.2. Results of the first research question

The source code of the scraper developed during the research is accessible in the GitHub repository of the dissertation at https://github.com/gneusch/JobPostingScraper. I started data collection January 16th, 2019 from the indeed.co.uk website and in a year, I have collected almost 400 thousand job listings.

In chapter 5.2 of the thesis I give an exhaustive list of data I wish to store, their source as well as the sort of information requirement they satisfy. Independently of the implementation, in order to illustrate the relations among the data I present a relational schema of the data that is relevant from the problem's point of view (Figure 1).

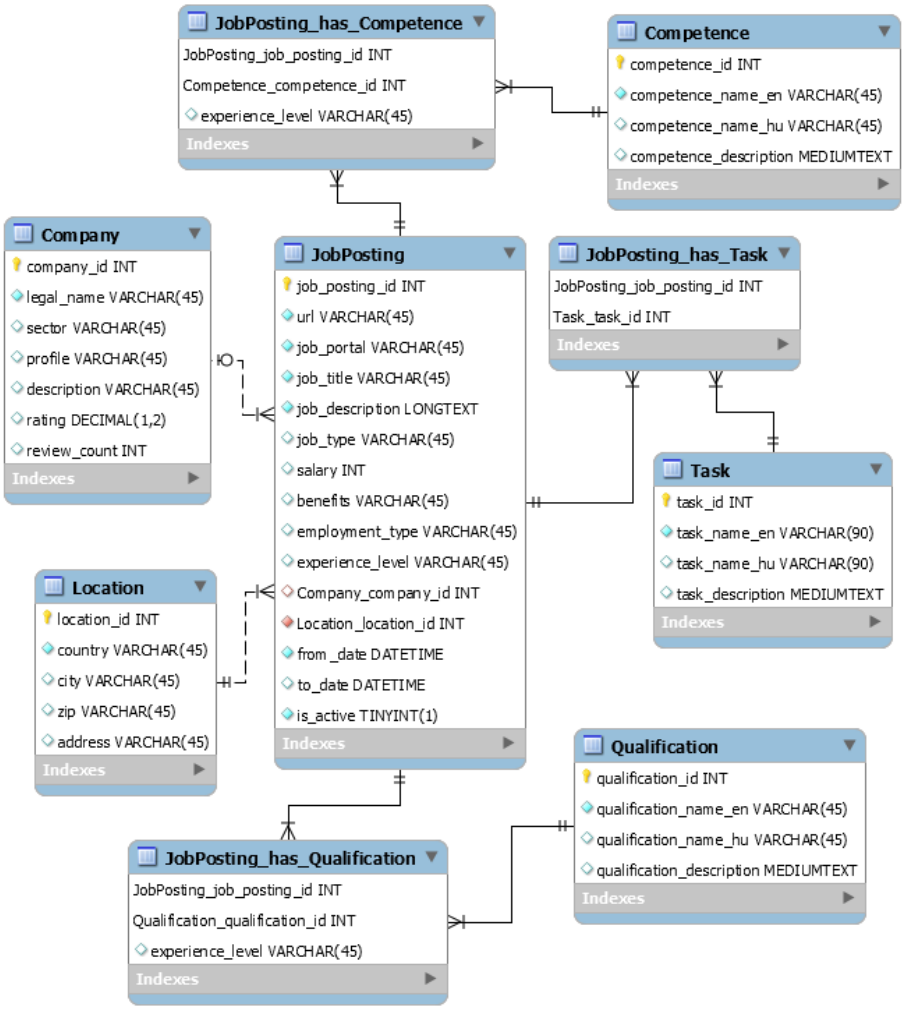

Figure 1.: The possible logical data model of the data warehouse

The most important selection and evaluation criteria of the data storage technology are the following:

- As a result of the *scraping* process the data are available in *JSON* format, structured along different attributes.
- However, numerous important information needs to be extracted from the description of job listings using further processing. During this task storing data in a staging area is necessary.
- I wish to store the job listings and their descriptions in their unprocessed form, so they would be available in the future if need be. I also would like to have the other outside systems' data to be available in the database to simplify and accelerate processing as well as to guarantee continuity.

- The most important requirement regarding the speed of the system is that user queries could be answered in real time that is why the extraction and storage of required competencies and jobs, etc. are necessary and for storage in this case a relational database is more suitable.
- It is important that the system would be well scalable and guarantee high availability. Strong consistency however is not expected.

In light of the previous, taking the advantages and disadvantages of the presented technologies into consideration, and the most important goal to answer users' information requirements as fast as possible, I would create a solution based on a hybrid architecture, the main elements of which are shown in Figure 2. This kind of hybrid approach, when we build a database architecture that utilizes the most suitable, even different solutions with regards to the given problem, is referenced as polyglot persistency by Sadalage and Fowler (2012).

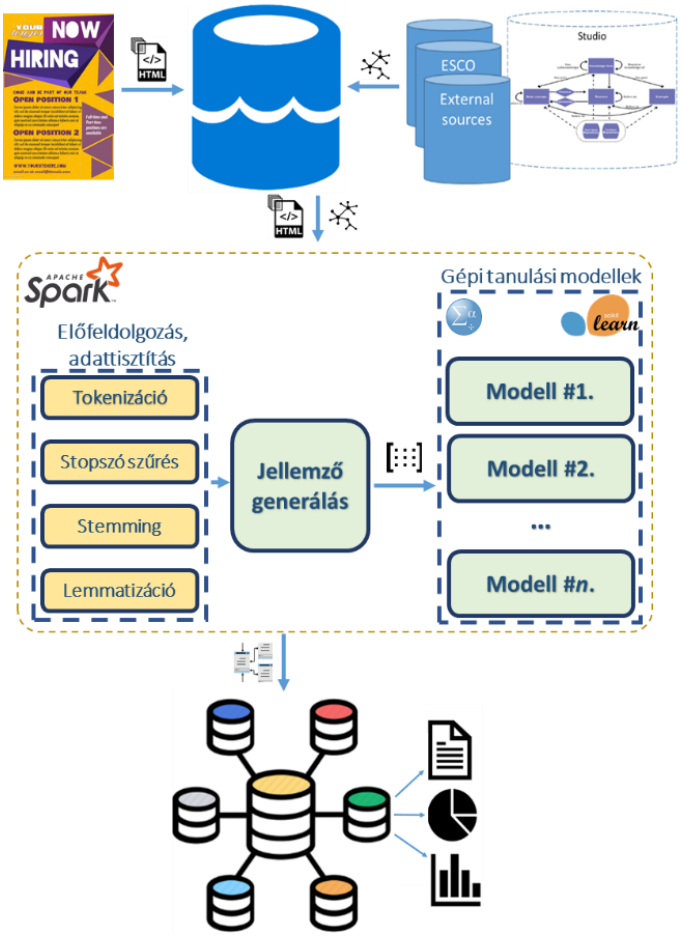

Figure 2. Hybrid storage solution's schematic architecture

In this architecture the collected job listings and the information of outside systems are first placed into a document database. So, there is no need to previously transform the data collected by the scraper, they can be stored as soon as they are downloaded, in their semi-structured form they get through the scraping process. Beyond providing a "staging area" another advantage of this design is that it is characterized by great flexibility, that is, if the structure of the received data changes or we wish to include new kinds of data in processing, etc. we do not have to make changes at this point of the framework. We can save into this document database data from outside systems or data we do not wish to use at the given moment, or do not fit into the schema to be created but might be useful later.

The document database however is not suitable to satisfy the possible users' requirements as fast as possible, among other things, because architecturally they are equipped to manage aggregates instead of slicing and dicing data along different dimensions; also, the processing of job descriptions (stored in text form in the postings) is expensive and takes time therefore its results are worth storing.

Consequently, I plan to do the linguistic processing of job descriptions and the extraction of the necessary information using machine learning algorithms in an "integrational" layer with a *Spark* application. Such design of the architecture is beneficial because it allows for further flexibility. For example data processing can be done any time and it can also be re-run if there is a change in the models or in the scope of the data used (for instance we include a new competence dictionary into the process or we "propagate back" expressions identified as competence during a previous iteration of the process). Not only is it a breeze to choose the algorithm used for extracting information but the used implementation as well. As I explain in the dissertation I built and tested my machine learning models in two different tools, in *IBM SPSS Statistics* and in the *Scikit-learn* software package (6.1.4 and 7.1.1.2 sub-chapters). Both resulting models are easy to integrate into a parallelized *Spark* application working on the data of a distributed document database.

After this the extracted information, now in structured form, can be loaded into the relational schema shown in Figure 1. using simple *SQL insert* and *update* commands. However, I would leave out the largest data element (as opposed to Figure 1.), job description, because I wish to store it long term in the suggested *NoSQL* database anyway

and I would like to avoid redundant storage and the need for horizontal scaling of the data warehouse.

So, with this architecture it can be guaranteed that each query can be answered very quickly based on only *SQL* operations and at the same time raw documents can be stored in the document database. This way, it is also ensured that if the process or the models need changing or new information is needed from the archived job descriptions, all data is available.

2.3. Methodology of research related to the second research question

I defined the most important goal of the data warehouse that it would provide executives with useful information about the job listings' competence content, about its evolution in time and space. However, these competencies are not available in structured form in job listings, but this information must be extracted from job descriptions using techniques related to natural language processing and machine learning. Thus, in accordance with the second research question, in chapter 6. of the dissertation I present such methods that can be potentially useful to identify "quality phrases", in this case competency candidates, in the unstructured text.

- In the beginning of the chapter, I defined the length of n-grams (sequence of tokens of length n) – used later in processing – on a statistical basis.
- I presented why models based on term- and document frequency (tf, df) or tf-idf (term frequency - inverse document frequency) values are unable or have a strongly limited ability to identify competence candidates.
- I have detailed the possibilities and the limitations of using a competence dictionary, built on the content of outside ontologies, for the extraction of information bearing terms.
- In the last block of the chapter, I gave a detailed presentation of 10 metrics, (Jaccard index, Sørensen-Dice- and cosine distance, etc.) that can be used for the definition of the similarity and distance of terms. After this I quantified the similarity between the elements of the competence dictionary and the n-grams generated from the descriptions of job listings.
- Using the results from the preceding step as feature variables in a logit model, I was trying to find the answer whether it is possible to create a model that can

quantify if a term can be accepted as a competence candidate. I used the *IBM SPSS Statistics* software package to build the model. Since I have done the testing of the model parameters on the teaching set, and there was no danger of data-peeking, that is why I did not generate a separate dataset used for validation. I used the *Backward Wald* method for feature selection, and I set the *cut value* in *0,4*.

2.4. Results of the second research question

The global indicators of the accepted model are adequate. Based on the Omnibus test the alternative hypothesis can be accepted on every significance level, that is, it is guaranteed that there is a variable in the model whose coefficient is significant. The pseudo R^2 measures show medium determination. Based on Cox and Snell's measure feature variables define the chance of a term identifying a real competence element in 45% while based on Nagelkerke R^2 index the determination is 78,8%. However, the direct interpretation of these indicators can be misleading, because they only state "by how many percent the log *likelihood* value that belongs to the model that contains constants only, was decreased." (Kovács, 2014; Fliszár *et al.*, 2016, p. 46). For the examination of the model fit the literature mostly recommends the Hosmer-Lemeshow test. During this, the observed values and predicted probabilities are divided into deciles and we examine the hypothesis whether the number of events is the same as the prediction in reference for each decile. This hypothesis is acceptable for the present model.

The *logit* model presented in the dissertation performed adequately on the test data; recall was 85% while precision was 71,9%. Since manual intervention to the process is inevitable in the beginning, that is, before we would accept these competence candidates as real competencies, a professional has to review the results; this way the model is acceptable as one that can provide useful information and carries added value. I accept the model from the point of view that it gives a satisfactory answer to my second research question and it is suitable to identify a significant proportion of the competencies that appear explicitly.

In the summary of the chapter, I elaborate on the directions in which I intend to carry on with the research in the future. The most important of these is the decrease of noise in the of job description by more thorough pre-processing to increase the precision of the model.

Development has taken place in this direction in the meantime, that have already been utilized during the experiments for the examination of the third research question. The other important direction that I want to take with the research in the future is the Hungarian adaptation of the model so that in the long run the framework would be applicable also for the analysis and regional comparison etc. of the national competence demand.

2.5. Methodology of research related to the third research question

In the third large block of the research, about the third research question, in Chapter 7 I examined the methods and applications that I deemed suitable for the identification of the competencies covertly present in job listings. The most important direction I analyzed in detail whether by using the content of the ESCO and the O*NET ontologies I can identify the jobs related to the positions advertised in the postings.

- To do this, I developed a method based on regular expressions and simple rules to identify the job in the title of a posting.
- Furthermore, I gave a detailed presentation of a classification model based on a decision tree and using it I attempted to match job listings – based on their title – to the job labels of the used ontologies. As decision criteria for this task the lexicographical and "quasi-semantic" similarity values of the job labels of the used ontologies and the titles of postings – cleared during the pre-processing steps – was used.
    - I used the *Scikit-learn* software package to build the model.
    - Due to the imbalanced dataset, considering that decision tree algorithms tend to create faulty models biased towards the dominant classes, I wanted to counterbalance this problem during my experiments, by the oversampling of observations that fell into the minority classes. I examined the performance of a simple random substitution based- and the *SMOTE* oversampling algorithm.
    - Also because of the imbalance of the data I generated the teaching and test sets using stratified sampling.
    - Due to the relatively low number of observations, instead of using separate teaching, validating and test sets, I chose the technique of cross-validation

for hyperparameter tuning and testing as well as for the estimation of the different models' prediction errors.

  - In order to avoid overfitting – that is, excessive over-learning for the teaching set and its idiosyncrasies and faults – and to find the most precise model, I tried to maximize the accuracy of the model along the different values of depth, the effective α parameter (of cost complexity pruning) and the minimal number of samples per leaf.

- Using the *t-SNE* method, I verified the results of Csepregi (2020) that says that clusters (subject groups) of job listings cannot be identified or only to a very limited degree and low efficiency based on the *tf-idf* matrix generated from their description.
- In closing Chapter 7 of the dissertation, I presented methods that can be used to explore covert requirements through the relations of competencies that appear explicitly. I have not examined these directions in detail in the dissertation, but I plan to further investigate them in depth in my further research.

2.6. Results of the third research question

In connection to the third research question I examined the possibilities of the identification of covert or latent competencies that do not appear explicitly in job listings. I have outlined three major conceptional directions. The first one was to identify the relevant latent competencies in the ontologies through the identification of the job related to the posting.

- Using the method built on regular expressions and simple rules I managed to link 23,7 % of the job listings in October 2019 to a job and I did so correctly in 97,5% of the cases.
- During the experiments with decision trees the model accepted as best had a 58% precision and 68% recall on the test set, which is not exceptional, but it is competitive with the models found in the literature created for similar tasks. For instance, Amato and co-authors (2015) used several methods to examine the identifiability of jobs in the title of job listings. Even though the authors did not give an in-depth, detailed description regarding the use of methods, but based on the published results it can be said that recall and precision, in case of LDA (*Latent*

Dirichlet Allocation) was around 50% while in the case of other methods (Linear Support Vector Machine, Perceptron Classifier) it was between 25-35%.

I want to repeat the experiment in the future, teaching the model with a larger dataset and using other methods, because, based on the jobs identified this way the competencies related to them in the ontologies are also acceptable as those that are implicitly relevant in reference to the posting. I accepted this direction from the point of view of the third research question as a possible solution that is, implicit competencies can be explored by identifying the jobs indicated in the listings, but the efficiency of this identification process has to be improved in the future.

However, I rejected the second examined direction, that listings can be grouped based on their description; and overlapping competence content, or some sort of job groups can be concluded based on these groups. I made this decision based on Csepregi (2020) and my own experiment done using the *t-SNE* method.

In the last part of Chapter 7 I presented the third possible direction for the exploration of latent competencies using the ontology relations of explicitly appearing competencies. Based on the results I arrived at in my master's thesis I can also accept this direction as one that is suitable for the identification of latent competencies as I proposed it in my third research question. Consequently, I wish to improve this area as well in future research.

3. References

Amato, F., Boselli, R., Cesarini, M., Mercorio, F., Mezzanzanica, M., Moscato, V., Persia, F., *et al.* (2015), "Classification of Web Job Advertisements: A Case Study", *23rd Italian Symposium on Advanced Database Systems (SEBD 2015)*, presented at the 23rd Italian Symposium on Advanced Database Systems (SEBD 2015) (Gaeta, 14/06/2015 - 17/06/2015), Curran, Gaeta, Italy, pp. 144–151, available at: http://hdl.handle.net/10863/10468.

Beck, M. and Libert, B. (2017), "The Rise of AI Makes Emotional Intelligence More Important", *Harvard Business Review*, 15 February, available at: https://hbr.org/2017/02/the-rise-of-ai-makes-emotional-intelligence-more-important (accessed 4 May 2019).

Bodon, F. (2010), "Adatbányászati algoritmusok", Dr. Bodon Ferenc, available at: http://www.cs.bme.hu/~bodon/magyar/adatbanyaszat/tanulmany/adatbanyaszat.pdf (accessed 1 October 2019).

Bodon, F. and Buza, K. (2013), "Adatbányászat", Elektronikus tananyag, available at: http://www.cs.bme.hu/~buza/pdfs/adatbanyaszat-cover.pdf (accessed 20 December 2020).

Chang, H.-C., Wang, C.-Y. and Hawamdeh, S. (2018), "Emerging trends in data analytics and knowledge management job market: extending KSA framework", *Journal of Knowledge Management*, Vol. 23 No. 4, pp. 664–686, available at: https://doi.org/10.1108/JKM-02-2018-0088.

Choi, S., Cha, S. and Tappert, C.C. (2010), "A Survey of Binary Similarity and Distance Measures", *Journal of Systemics, Cybernetics and Informatics*, Vol. 8 No. 1, pp. 43–48.

Csepregi D. (2020), *Első Lépés az Automatizált Oktatásfejlesztés Felé: Hogyan tehet szert előnyre szövegelemzési eszközökkel a Corvinus Gazdaságinformatikus képzése?*, TDK dolgozat, Budapesti Corvinus Egyetem, Gazdálkodástudományi kar, Budapest, available at: http://publikaciok.lib.uni-corvinus.hu/publikus/tdk/csepregi_d_2020a.pdf (accessed 27 December 2020).

Fajszi, B. and Cser, L. (2004), *Üzleti Tudás Az Adatok Mélyén*, Budapesti Műszaki és Gazdaságtudományi Egyetem.

Fajszi, B., Cser, L. and Fehér, T. (2010), *Üzleti Haszon Az Adatok Mélyén*, Alinea Kiadó - IQSYS Informatikai és Tanácsadó Zrt.

Fazekas, K. (2017), *Nem Kognitív Készségek Kereslete És Kínálata a Munkaerőpiacon*, Institute of Economics, Centre for Economic and Regional Studies, Hungarian Academy of Sciences, Budapest, available at: http://www.mtakti.hu/wp-content/uploads/2017/11/FK-BWP1709-jav-OE_FKjav.pdf (accessed 20 July 2019).

Fliszár, V., Kovács, E., Szepesváry, L. and Szüle, B. (2016), *Többváltozós Adatelemzési Számítások*, Budapesti Corvinus Egyetem, available at: http://unipub.lib.uni-corvinus.hu/2438/.

Gábor, A., Kő, A., Szabó, Z. and Fehér, P. (2016), "Corporate Knowledge Discovery and Organizational Learning: The Role, Importance, and Application of Semantic Business Process Management—The ProKEX Case", in Gábor, A. and Kő, A. (Eds.), *Corporate Knowledge Discovery and Organizational Learning*, Springer International Publishing, pp. 1–31, available at: https://doi.org/10.1007/978-3-319-28917-5_1.

Gajdos, S. (2019), *Adatbázisok*, A 2015. évi kiadás negyedik javított utánnyomása., BME, Budapest, Magyarország.

Gruber, T.R. (1993), “A Translation Approach to Portable Ontology Specifications”, *Knowledge Acquisition*, Vol. 5(2), pp. 199–220, available at: http://tomgruber.org/writing/ontolingua-kaj-1993.htm.

Grundke, R., Marcolin, L., Nguyen, T.L.B. and Squicciarini, M. (2018), “Which skills for the digital era?”, *OECD Science, Technology and Industry Working Papers*, available at: https://doi.org/10.1787/9a9479b5-en.

Gugnani, A. and Misra, H. (2020), “Implicit Skills Extraction Using Document Embedding and Its Use in Job Recommendation”, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34 No. 08, pp. 13286–13293, available at: https://doi.org/10.1609/aaai.v34i08.7038.

Kovács, E. (2014), *Többváltozós Adatelemzés*, Typotex Kiadó.

Manning, C.D., Raghavan, P. and Schütze, H. (2009), *Introduction to Information Retrieval*, Online edition., Cambridge University Press, New York.

Nasir, S.A.M., Yaacob, W.F.W. and Aziz, W.A.H.W. (2020), “Analysing Online Vacancy and Skills Demand using Text Mining”, *Journal of Physics: Conference Series*, IOP Publishing, Vol. 1496, p. 12, available at: https://doi.org/10.1088/1742-6596/1496/1/012011.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., *et al.* (2011), “Scikit-learn: Machine Learning in Python”, *Journal of Machine Learning Research*, Vol. 12, pp. 2825–2830, available at: https://dl.acm.org/doi/10.5555/1953048.2078195.

Pitukhin, E., Varfolomeyev, A. and Tulaeva, A. (2016), “JOB ADVERTISEMENTS ANALYSIS FOR CURRICULA MANAGEMENT: THE COMPETENCY APPROACH”, *ICERI2016 Proceedings*, presented at the 9th annual International

Conference of Education, Research and Innovation, IATED, Seville, Spain, pp. 2026–2035, available at: https://doi.org/10.21125/iceri.2016.1456.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J. and Manning, C.D. (2020), "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, presented at the ACL, Association for Computational Linguistics, Online, pp. 101–108, available at: https://doi.org/10.18653/v1/2020.acl-demos.14.

Russell, S. and Norvig, P. (2005), *Mesterséges Intelligencia Modern Megközelítésben*, Panem Kft., Budapest, available at: https://mialmanach.mit.bme.hu/aima/index (accessed 22 April 2017).

Sadalage, P.J. and Fowler, M. (2012), *NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence*, Addison-Wesley, available at: https://dl.acm.org/doi/book/10.5555/2381014.

Szabó Z. (2000), *A szervezeti információfeldolgozás strukturális és technológiai tényezőinek összerendelése*, phd, Budapesti Corvinus Egyetem, available at: http://phd.lib.uni-corvinus.hu/212/ (accessed 24 January 2021).

Varga, K. (2014), *A Szemantikus Folyamatmenedzsment Hasznosítási Lehetősége Az Üzleti Folyamatok Tudásalapú Fejlesztésében*, PhD thesis, Budapesti Corvinus Egyetem, available at: http://phd.lib.uni-corvinus.hu/818/ (accessed 13 September 2019).

Vas, R.F. (2007), *Tudásfelmérést Támogató Oktatási Ontológia Szerepe És Alkalmazási Lehetőségei*, PhD thesis, Budapesti Corvinus Egyetem, available at: http://phd.lib.uni-corvinus.hu/258/.

Waldrop, M.M. (2016), “The chips are down for Moore’s law”, *Nature News*, Vol. 530 No. 7589, pp. 144–147, available at: https://doi.org/10.1038/530144a.

Wowczko, I.A. (2015), “Skills and Vacancy Analysis with Data Mining Techniques”, *Informatics*, Vol. 2 No. 4, pp. 31–49, available at: https://doi.org/10.3390/informatics2040031.

Zhao, M., Javed, F., Jacob, F. and McNair, M. (2015), “SKILL: a system for skill identification and normalization”, *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI Press, Austin, Texas, pp. 4012–4017, available at: https://dl.acm.org/doi/10.5555/2888116.2888273.

4. Publications

Peer reviewed academic journals

Neusch, G. (2014), “Domain Ontology Tailoring Based on Business Processes in the Frame of the ProKEX Project”, *SEFBIS Journal*, Vol. I No. XI, pp. 51–59.

Szabó, I. and Neusch, G. (2015), “Dynamic Skill Gap Analysis Using Ontology Matching”, in Kő, A. and Francesconi, E. (Eds.), *Electronic Government and the Information Systems Perspective*, Vol. 9265, Springer International Publishing, pp. 231–242, available at: http://dx.doi.org/10.1007/978-3-319-22389-6_17.

Beel, J., Carevic, Z., Schaible, J. and Neusch, G. (2017), “RARD: The Related-Article Recommendation Dataset”, *D-Lib Magazine*, Vol. 23 No. 7/8, available at: https://doi.org/10.1045/july2017-beel.

Szabó, I., Neusch G., Vas R. (2021, megjelenés alatt), „Design Thinking based Ontology Development for Robo-Advisors”, *Proceedings of the 20th International*

Conference Intelligent Systems Design and Applications (ISDA 2020), Advances in Intelligent Systems and Computing, Springer Verlag

Studies published in peer-reviewed conference volume

Neusch, G. and Gábor, A. (2014), "ProKEX – Integrated Platform for Process-Based Knowledge Extraction", in Gómez Chova, L., López Martínez, A. and Candel Torres, I. (Eds.), *ICERI2014 Proceedings*, presented at the 7th International Conference on Education Research and Innovation, IATED Academy, Seville, Spain.

Castello, V., Mahajan, L., Flores, E., Gabor, M., Neusch, G., Szabo, I., Guerrero, J., *et al.* (2014), "THE SKILL MATCH CHALLENGE. EVIDENCES FROM THE SMART PROJECT", in Gómez Chova, L., López Martínez, A. and Candel Torres, I. (Eds.), *ICERI2014 Proceedings*, IATED Academy.

Weber, C., Neusch, G. and Vas, R. (2016), "Studio: A Domain Ontology Based Solution for Knowledge Discovery in Learning and Assessment", *Proceedings of the 2016 AIS SIGED International Conference on Information Systems Education and Research*, pp. 1-13., available at: https://aisel.aisnet.org/siged2016/12.

Neusch, G. (2016), "Ontology Tailoring for Job Role Knowledge", in Gábor, A. and Kő, A. (Eds.), *Corporate Knowledge Discovery and Organizational Learning*, Springer International Publishing, pp. 105–130, available at: https://doi.org/10.1007/978-3-319-28917-5_5.

Other professional achievements

Neusch, G., 2015. 'Studio-User Help' saját számítógépi programalkotás, Szellemi Tulajdon Nemzeti Hivatala, Önkéntes műnyilvántartásba vételi szám: 003871