Ontology-guided Job Market Demand Analysis

A Cross-Sectional Study for the Data Science field

Elisa Margareth Sibarani
University of Bonn, Germany &
Institut Teknologi Del, Indonesia
sibarani@cs.uni-bonn.de

Simon Scerri
University of Bonn / Fraunhofer IAIS
scerri@cs.uni-bonn.de

Camilo Morales
Douglas GmbH
c.morales@douglas.de

Sören Auer
Leibniz Universität Hannover /
TIB Information Center
soeren.auer@tib.eu

Diego Collarana
University of Bonn / Fraunhofer IAIS
collaran@cs.uni-bonn.de

ABSTRACT

The rapid changes in the job market, including a continuous year-on-year increase in new skills in sectors like information technology, has resulted in new challenges for job seekers and educators alike. The former feel less informed about which skills they should acquire to raise their competitiveness, whereas the latter are inadequately prepared to offer courses that meet the expectations by fast-evolving sectors like data science. In this paper, we describe efforts to obtain job demand data and employ a information extraction method guided by a purposely-designed vocabulary to identify skills requested by the job vacancies. The Ontology-based Information Extraction (OBIE) method employed relies on the *Skills and Recruitment Ontology* (SARO), which we developed to represent job postings in the context of skills and competencies needed to fill a job role. Skill demand by employers is then abstracted using co-word analysis based on a set of skill keywords and their co-occurrences in the job posts. This method reveals the technical skills in demand together with their structure for revealing significant linkages. In an evaluation, the performance of the OBIE method for automatic skill annotation is estimated (strict F-measure) at 79%, which is satisfactory given that human inter-annotator agreement was found to be automatic keyword indexing with an overall strict F-measure at 94%. In a secondary study, sample skill maps generated from the matrix of co-occurrences and correlation are presented and discussed as proof-of-concept, highlighting the potential of using the extracted OBIE data for more advanced analysis that we plan as future work, including time series analysis.

CCS CONCEPTS

•Information systems →Web Ontology Language (OWL); Content analysis and feature selection; Information extraction; Clustering and classification;

Semantics2017, Amsterdam, Netherlands

DOI: 10.1145/3132218.3132228

KEYWORDS

Co-word Analysis, Ontology-based Information Extraction, Automatic Keyword Indexing, Job Adverts

ACM Reference format:
Elisa Margareth Sibarani, Simon Scerri, Camilo Morales, Sören Auer, and Diego Collarana. 2017. Ontology-guided Job Market Demand Analysis. In *Proceedings of Semantics2017, Amsterdam, Netherlands, September 11–14, 2017,* 8 pages.
DOI: 10.1145/3132218.3132228

1 INTRODUCTION

In the past twenty years, the use of the Web has increased dramatically and as a consequence, job advertisements are now mainly published electronically online [15]. In addition, due to the digitization of society and economy, the job markets are consistently and profoundly changing. New job profiles are emerging and corresponding skills are in high-demand, while formerly important skills become irrelevant. As a result, it is of paramount importance to give job seekers and education providers a timely and comprehensive overview of the current situation on the job market in terms of required skills, competences, and technologies.

The aim of this study is to identify the most needed technical skills that are important for a data scientist's work by analyzing job advertisements. We performed a cross-sectional analysis which focuses on a snapshot of the demand of the current job market. The first target user group is job seekers and applicants, with the purpose to supply them with the skills in demand. Secondly, we targeted educators and training providers by helping them to determine which courses are in high demand and should be added to their curricula.

Our study portrays the skills in demand at a point in time by utilizing co-word analysis, as a quantitative method, that is well known to identify and structure relationships among concepts. The rationale is that when two skills often appear together in job adverts, there is a high chance that both skills are strongly related to the job role to which they refer. The analysis is done based on the occurrence and co-occurrence frequency of skill pairs in the job adverts. Its methodological foundation is that the co-occurrence of keywords describes the contents of the documents [1]. To extract the keywords prior to the co-word analysis, we chose the *Ontology-based Information Extraction* (OBIE) approach to annotate,

extract, and classify known entities within job adverts. The OBIE pipeline exploits an ontology by extracting pre-defined concepts and annotating text using concepts defined in the ontology.

For reaching good coverage, we introduce the *Skills and Recruitment Ontology* (SARO) ontology which extends the 'Job Posting' taxonomy from schema.org[1] and the 'Skills and Competences' concepts from the ESCO ontology[2]. Currently, the SARO ontology comprehensively covers IT and data science related skills, since this domain is the focus of our study.

Previous research has established that co-word analysis reduces and projects data into a specific visual representation while maintaining essential information [1, 5, 7, 17]. There is a growing body of literature that recognizes the potential of utilizing co-word analysis for investigating job advertisements, for instance, library and information science (LIS) job adverts [12, 13] and information systems (IS) job adverts [9, 10]. A review of research studies in LIS that uses job adverts to analyze and track changes to job skills, highlights key insights for researchers [6]. The goal of the review is to support the improvement of the method and enable future researchers to conduct more rigorous research and methodology of job adverts.

To the best of our knowledge there has been no detailed investigation that offers automated or semi-automated IE methods and co-word analysis targeting job advertisements. Hence, our contributions are in particular:

- An ontology called SARO[3] based on schema.org and ESCO ontologies, which contains more than 800 concepts representing the domain knowledge related to job posting attributes, skills, and qualifications.
- A software architecture and implementation of the OBIE method to process job adverts by extracting skills and storing the results as an RDF knowledge base.
- A comprehensive co-word analysis of 872 job adverts to determine skill demand based on a co-word matrix that is built by querying and analyzing the RDF knowledge base.

The evaluation of our OBIE method (section 4) indicates the feasibility of the approach with an overall strict and lenient F-measure of 0.79 and 0.83, respectively. This result represents a solid foundation for further improvements, e.g. the annotation of `SkillTool`, `SkillProduct`, and `SkillTopic`, by implementing additional rules to identify skills employing the JAPE Grammar in GATE Embedded. Also, the co-word analysis based on the OBIE method and guided by SARO ontology, can now be used to perform time-series analysis of job skills for trend identification. Due to its generic nature, the proposed OBIE method can be also used for other domains that require ontology-based keyword extraction and the SARO ontology can be reused for other tasks related to semantic job profile management.

The article is structured as follows: In section 2 we give a detailed overview on related research. In section 3, we present the implementation of the proposed system. Next, section 4 details the results and evaluation of the information extraction on a comprehensive corpus of job adverts. The result of the co-word analysis is explained in detail in section 5. Finally, section 6 concludes with an outlook on future work.

2 RELATED WORK

There exist, in the field of IT related needs assessments, publications that have published similar and relevant findings to our current research. For example, Todd et al. [16] compiled skill requirements over time for programmers, systems analysts, and managers by manually collecting, classifying, counting, and building an index of keywords from a pilot sample of 200 ads. Litecky et al. [11] performed a cluster analysis in two phases, hierarchical agglomerative clustering and k-means, to classify 209k job adverts in computing, into 20 clusters of job definitions based on the skills specified.

Surakka [15] conducted a trend analysis from the period of 1990-2004 and a cross-sectional analysis from year 2004, to identify technical skills for the software developer position from American job adverts. All of the job adverts were read and coded manually and a quantitative content analysis was conducted by simply calculating the frequencies of different phrases.

Marion et al. [12] explored 395 library and information science (LIS) job adverts in Australia and the US from August to October 2004, to define the technical skills together with their structure. They created a dictionary of 18 broad categories based on the counts of most frequently mentioned terms relevant to the study. Further content analysis was implemented using the software package SimStat/WordStat[4] to identify and convert the frequency count to a matrix of co-occurrence similarity (correlation) values. Finally, the structure of the correlation matrix was explored using two multivariate techniques, (cluster analysis and multidimensional scaling), which are part of the SimStat/WordStat software package.

Sanchez-Cuadrado et al. [13] examined the LIS professional skills from 1K job offers between 2006 and 2008 collected from a Spanish employment agency website. Similar with [12], this research used software package WordStat plus QDA Miner for the content analysis. The document indexing process was supported by a new thesaurus, containing the academic profiles of professionals in LIS which comprises 479 skills' terminology in Spanish. However, the paper did not provide any detailed information regarding their knowledge structure therefore, it is impossible to reuse it for the purpose of the present study.

Kennan et al. [10] studied 400 information systems (IS) job adverts between July 12 and September 13, 2006, in order to understand the skills and competencies demanded of early career IS graduates in Australia. Their critical contribution was to build a dictionary of 17 categories by using the SimStat/WordStat, measure the co-occurrence, and implement the cluster analysis, which was also included in the SimStat/WordStat software package.

Hu et al. [8] revealed the intellectual structure of LIS in China, utilizing co-word analysis by extracting keywords from the Chinese Journal Full-Text Database between 2008 and 2012. By manually filtering duplicated and irrelevant articles, they gained 80k keywords and created further processes, such as standardizing using "Chinese classified thesaurus", merging or altering terminology, and filtering general terms to get a final 181 keywords. Further clustering analysis was used with the aid of SPSS19.0 and the analysis of the

[1] http://schema.org/JobPosting
[2] https://ec.europa.eu/esco/resources/data/static/model/html/model.xhtml
[3] http://vocol.iais.fraunhofer.de/saro/
[4] https://provalisresearch.com/

network characteristics of the co-word matrix was conducted using Ucinet6.0.

Wowczko [19] analyzed skill needs from online IT vacancies during 2014 with RapidMiner and R. The pre-processing steps were done manually, resulting in two term-document matrices based on term frequency for job titles and descriptions with 58 and 2503 most frequent terms, respectively. By implementing K-Nearest Neighbors to cluster the frequent keywords into seven occupations, a matrix of bi-grams (two consecutive words) was built for each group, and R's wordcloud package was used to visualize the top 20 bi-grams for each occupations.

In 2012, Harper [6] reviewed the research methodology of 70 researches in library and information science (LIS) that used job adverts as a data source to analyze and track changes to job skills and the employment market over time. His study concluded that only the minority of these studies used automatic text analysis (3 out of 70) and inferential statistics (18 out of 70), despite the long history and large volume of studies examining job adverts in LIS.

A survey of OBIE applications was provided in [18] and a more recent survey in [14]. As OBIE has recently emerged as a subfield of information extraction (IE), an ontology - which provides formal and explicit specifications of conceptualizations - plays a crucial role in the IE process. Because of the use of ontology, this field is related to knowledge representation and has the potential to assist the development of the Semantic Web [18].

2.1 Reflection on the research

There has been relatively little research in the co-word analysis exploring the OBIE technique for the extraction of keywords. By considering the assessment points listed by [6], the following characteristics are an adaptation for the improvement of the present research in comparison with the previous job advertisement research:

to provide SARO ontology. – the *Skills and Recruitment Ontology*[5] is not just a dictionary of categories, but a knowledge representation that serves as a reusable model for other similar use case and domain;

to provide OBIE methodology. – an automatic text analysis to extract keywords guided by SARO ontology. OBIE is a critical step in our study to cope with the "indexer effect" of co-word analysis [7] which: (1) reduces the time-consuming and eliminates biased or inconsistent coding of manual indexing; (2) reduces, if not erases the delay between the publishing of the job adverts and the moment it is available for analysis. The text analysis software that is mostly used by the past research, SimStat/WordStat and QDAMiner, and another text analysis tool which, was suggested by [6], known as NVivo[6], are in fact not a freeware or open-source software; and

to utilize co-word analysis. – implement inferential statistical analysis with co-word analysis to reveal the skill demands together with their internal and external correlations with other skills in a different sub-network. Within the last three decades, this technique has been implemented by several research groups, who successfully discovered its power for knowledge discovery in structured or unstructured data [7]. In particular, we identified the Paris/Keele co-word method [17], [7] due to its reported superior adequacy in analyzing large and heterogeneous datasets. In contrast to similar methods it does not use complicated statistical techniques for assigning words to clusters. As a result, it generates a cluster structure which is, even for large numbers of keywords, simpler and easier to present in forms which non-specialists in the technique will find comprehensible [17].

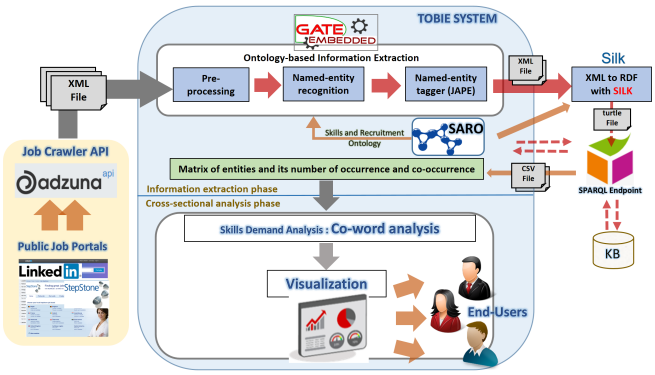

Figure 1: Architecture of our system implementation involving crawling, pre-processing, NER, linking and co-word analysis.

3 CO-WORD ANALYSIS FOR JOB ADVERTS

Employing co-word analysis as basis for our methodology, we use the Paris/Keele co-word method to reveal the skills' demand for data scientists in the UK and the description of the structure of skills embodied in the job postings. Figure 1 depicts our approach and comprises two main phases: (1) the OBIE pipeline guided by the SARO ontology and (2) the co-word analysis. Our implementation is developed in Java environment and will serve as a basis for our future work on time-series analysis. Its goal is to support the longitudinal study of skills demand, because co-word analysis has the potential of effectively revealing patterns and trends in a specific discipline [5].

In co-word analysis for job adverts, the presence of many co-occurrences around the pair of skill words means stronger correlation in both, which points to a suggestion that two skills are related to a specific job role.

3.1 Ontology development and enrichment

We perform a survey of existing vocabularies related to employment in order to elaborate the ontological model. Our SARO ontology is built as the extension of mainly two relevant models which are: (1) the European Skills, Competences, Qualifications and Occupations (ESCO) ontology, which focuses on the labor market and its skills and qualifications [4]; and (2) Schema.org[7], which describes job openings in organizations, offering properties such as `datePosted`, `hiringOrganization`, and `jobLocation`.

[5] http://vocol.iais.fraunhofer.de/saro/
[6] http://www.qsrinternational.com/
[7] http://schema.org/

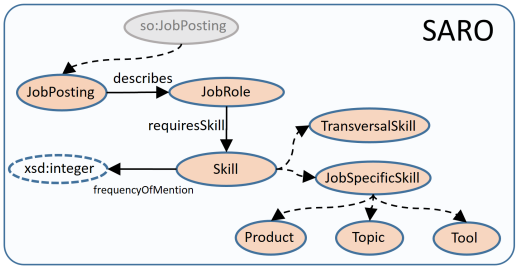

Figure 2: Upper level view of the SARO ontology, showing the `Skill` and `JobPosting` concepts and the relationships between these.

SARO aims to enable optimal analysis and reuse in a variety of contexts related to tasks, which required defining and correctly interpreting job postings in the context of skills and competencies. This ontology includes concepts, relationships, and instances as the formalization in the context of skills and competencies needed to fill a job role.

Figure 2 provides a top-level view of SARO, which is centered around the following two core concepts.

`saro:JobPosting.` – refers to a job advert listed by a specified `hiringOrganization`. It extends the `JobPosting`[8] concept in schema.org and defines essential attributes – including the `so:jobLocation`, `so:datePosted`, and `so:hiringOrganization`, in addition to the `saro:describes` to state the `saro:jobRole` of the `saro:JobPosting`.

`saro:Skill.` – a `saro:JobPosting` links to a set of inferred and explicitly specified `saro:Skills`, using the relation `saro:requiresSkill`. Another key relation is `saro:frequencyOfMention`, used to specify the number of mentions (occurrence) of a `saro:Skill` in a `saro:JobPosting`. `saro:Skill` extends `esco:Concept`, which categorizes skills (or competencies) as *job-specific* or *transversal* (cross-sector). SARO further extends this into:

(1) `saro:JobSpecificSkill`: representing technical skills related to a particular job role, further sub-classed into:
   (a) `Product`: competence using a particular product (e.g., "Hadoop").
   (b) `Topic`: capability in a domain- and/or role-specific topic required to achieve an observable result (e.g., "Data Analytics").
   (c) `Tool`: competence in the use of a tool specifically for carrying out technical tasks, e.g., a specific programming language (e.g., "Java", "Python") or database type (e.g., "NoSQL").
(2) `saro:TransversalSkill`: sector and occupation-independent skills foundational to personal development, often referred to as *soft* skills (e.g., "team-working").

Further, the necessary skills' terminology should be inserted to support the automatic identification of terms in the job advertisements. Overall, nearly 800 terms were manually added to the ontology based on the findings of the mentioned skills in the ads that were relevant to the study.

3.2 Ontology-based information extraction and pre-processing

In this study, the IE process builds the index list by processing a set of job adverts and extracts information regarding skills and related job posting attributes. A certain model that specifies the objective of the search (e.g. job title, location, skill, etc.) is required to guide this process. Therefore, we employ Ontology-based Information Extraction (OBIE) to identify the components of an ontology within text.

From the variety of available Natural Language Processing (NLP) tools supporting OBIE methods, the GATE framework which includes the ANNIE [3] IE system was chosen as the most appropriate for a number of reasons. Primarily, it enables the developer to implement a more flexible IE system by allowing the embedding of language processing functionality in diverse applications through GATE Embedded. Secondly, it supplies robust evaluation tools for NLP involving F-measure calculation based on a gold standard and inter-annotator agreement, thus decreasing the required effort.

Our ontology-based information extraction pipeline consists of linguistic annotation components, which accepts a set of job adverts and the SARO ontology as the input to guide the extraction process. It was built with the GATE Embedded framework and comprises of: (1) the linguistic analysis components (for pre-processing), (2) a named-entity recognizer based on the ontology, and (3) a named-entity tagger.

After OBIE successfully retrieved all important information regarding job posting and its skills, several tasks are performed: (1) converting the OBIE result from XML to RDF using *SILK*[9] framework, SILK allows us to easily transform the OBIE XML output into `JobPosting` and `Skill` concept instances adhering to the SARO Ontology; (2) loading all extracted RDF triples into the knowledge base; and (3) querying the set of triples to obtain the number of occurrences and co-occurrences through a SPARQL end-point and constructing a symmetric co-word matrix.

Further, the matrix is transformed into a correlation matrix, using an index to measure the strength of association between two keywords called, the equivalence index (e-coefficient) or strength ([7], [1]):

$$E_{ij} = (C_{ij})^2/(C_i \cdot C_j) \quad (1)$$

where, E_{ij} has a value between 0 and 1; C_{ij} is the number of job adverts in which the skill pair appears; C_i is the number of times that the keyword i is used for indexing a document from the document set; C_j is the number of times that the keyword j is used for indexing a document from the document set.

3.3 Co-word analysis

We used the Paris/Keele method for cluster analysis ([17], [7]). It rests upon the assumption that there is a cluster-type structure and employs an algorithm based on a threshold of the co-occurrence frequency and the number of links in one cluster (or sub-network).

[8] http://schema.org/JobPosting

[9] http://silkframework.org/

The process of constructing clusters (or sub-networks) is divided into two "passes". During Pass-1, the network is constructed by choosing the link that has the highest e-coefficient. These linked nodes become the starting points for the first pass-1 network. Other links and their corresponding nodes are added into the sub-network in the decreasing order of their e-coefficient based on a breadth-first search, until there are no more links that exceed the co-occurrence threshold, or a maximum pass-1 link limit is exceeded.

In Pass-2, each Pass-1 sub-network is extended by adding links that have the highest strength that exceed the co-occurrence threshold and both nodes of the link must be included in some Pass-1 sub-network. The Pass-2 algorithm will continue until no remaining link meets the co-occurrence threshold, or when the total link Pass-2 is met.

Subsequently, a keyword's structure/map could be drawn to visualize the resulted network, comprising sub-network(s) and all external associations between different sub-networks. The goal is to discover the central skills in a domain and depict their relationship to skills that occur less frequently, but are potentially growing areas.

Another map called strategic diagram is employed to illustrate the internal strength of the network ("local") and the degree of interaction with another network ("global") [7]. The goal is to reveal the status and simplified presentation of the current skills' demand of data scientists, and provide a stepping-stone for a dynamic analysis. For a given cluster, its density value is an index to measure its internal strength (local context) and is defined by the average (mean) of the e-coefficient (strength) values of the pass-1 links of the network. Next, the centrality value is to measure the strength of a cluster's interaction with other clusters (global context), which was calculated with the square root of the sum of the squares of all external link values (pass-2 links).

The strategic diagram consists of the horizontal and vertical axes, which represents centrality and density, and the origin of the graph is at the median of the respective axis values [7]. It is drawn by plotting centrality and density of each skills' network within a two-dimensional space divided into four quadrants.

To sum up, we can describe a network as follows [1, 7]: (1) Clusters of quadrant 1 are both central to the general network and internally coherent (they display a high degree of development). These clusters in some sense constitute the file's core and their position is strategical; (2) Clusters in quadrant 2 are weakly structured areas since they are weakly linked together, however individually, are linked strongly to specific skill areas throughout the network. In other words, attention to these areas appears to be under-developed, but it could potentially be of considerable significance to the entire skill demand network; (3) Clusters at quadrant 3 are internally well-structured and indicate that a group of employers still demand them. However, they are not central (call them peripheral) to the job role being sought; and (4) Clusters of quadrant 4 seem to be of only marginal interest to the job role in demand.

Our system receives five parameters: Pass-1Links, maxLinks, Co-occurThreshold, inputDatafile and outputDatafile. Respectively, these indicate the maximum number of Pass-1 links allowed, the maximum links which are allowed in total for one network, the minimum co-occurrence that is needed to create a link, the correlation matrix in csv format, and the output files in csv format. In particular, minimum co-occurrence is required to build the Pass-1 networks in order to prevent irrelevant or weak associations from dominating the network. Although two keywords may appear together more frequently in the dataset, they could still have lower association strength in comparison to two keywords which appear infrequently but always appear together in the dataset. Therefore, only keywords whose links fall above a chosen threshold are used to build clusters. This is the crucial difference from the normal classification algorithms.

Finally, by using Python, iGraph library to generate graphs, and Plotly library for the bubble plots, we visualize the co-word analysis result in two kinds of maps, keyword networks and strategic diagrams.

Table 1: Keywords ranked by frequency of occurrence

Keywords	Rank	No.Ads	% Ads
Analyst	1	773	88.6
Analysis	2	492	56.4
Development	3	378	43.3
Excel	4	372	42.6
Reporting	5	245	28.1
Marketing	6	216	24.7
Analytic	7	206	23.6
Design	8	200	22.9
SQL	9	198	22.7
Analytics	10	181	20.7
Finance	11	162	18.6
SAS	12	132	15.1
Database	13	120	13.7
Project Management	14	102	11.7
Assurance	15	100	11.5

4 USER STUDY AND EVALUATION

The objectives of our study are two-fold: i) to evaluate the adequacy of the OBIE method in extracting the relevant information and its performance compared to manual human extraction, and ii) to identify whether the proposed co-word analysis method has the potential to yield useful insights into complementary skills identified by job market demand observations. In contrast to the former exercise, which determined the F-measure of the OBIE method, the latter does not return a quantitative or qualitative scientific result. Instead, it offers a proof-of-concept of what the generated analysis can look like based on actual examples, ahead of the next stage of our research, which will investigate time-series analysis based on a series of so-generated skill networks.

4.1 Data Collection and pre-processing

For our study we rely on 872 job adverts between August to November 2015, crawled from Adzuna.com using the Adzuna API[10]. The execution of the SARO-guided OBIE method yields 184 keywords (skills) and attributes, ranging from 1 to 26 per advert. Around 20% of job adverts have at least 10 keywords, whereas 95% of job adverts had more than one keyword. Table 1 ranks the top keywords.

4.2 Evaluation of the OBIE method

In this experiment, we compute the precision, recall, and F-measure for extracting job post attributes and skills using the OBIE pipeline, and compare it to the results of with respect to the Inter-Annotator

[10] API: https://developer.adzuna.com/overview

Table 2: Strict IAA for Two Annotators

	Precision	Recall	F-measure
Organization	1.0	0.98	0.99
datePosted	1.0	1.0	1.0
jobLocation	0.98	0.96	0.97
jobRole	0.98	0.98	0.98
SkillProduct	0.90	0.89	0.89
SkillTool	1.0	0.94	0.97
SkillTopic	0.93	0.93	0.93
Summary	0.94	0.94	0.94

Table 3: F-measure Strict and Lenient for OBIE Evaluation

	Prec.	Rec.	F1-Strict	Prec.	Rec.	F1-Lenient
Organization	0.98	0.99	0.98	0.99	1.0	0.99
datePosted	1.0	0.98	0.99	1.0	0.98	0.99
jobLocation	0.98	0.97	0.98	1.0	0.99	0.99
jobRole	0.77	0.97	0.86	0.79	0.99	0.88
SkillProduct	0.83	0.57	0.68	0.90	0.62	0.73
SkillTool	0.52	0.68	0.59	0.53	0.70	0.60
SkillTopic	0.89	0.63	0.74	0.96	0.68	0.80
Overall	0.88	0.71	0.79	0.93	0.75	0.83

Agreement (IAA) achieved for a manually-annotated subset of job postings.

4.2.1 Gold Standard and Inter-Annotator Agreement. To determine the highest F-score that can realistically by expected for the OBIE pipeline, we instructed two annotators to independently annotate a random sample of 50 job postings. The annotators were familiar with job adverts and they were also introduced to the SARO concepts. Using GATE's inbuilt plugin, the strict IAA F-score is computed at 94%. The output, shown in Table 2, shows the results for seven annotation classes derived from SARO: `50 datePosted`, `50 jobLocation`, `50 jobRole`, `49 Organization`, `84 SkillProduct`, `17 SkillTool` and `482 SkillTopic`. The results show that whereas in some cases human agreement was full or close to complete, some types of annotations (e.g. SkillTopic: skills that are categorised neither as a tool, or product) are more subjective.

The IAA results were discussed with both annotators to gain an insight into their annotation subjectivity and identify a version with full agreement. Following this discussion, both annotators manually annotated two additional and different samples of 50 job posts. These were combined with the agreed-on sample to generate a gold standard of manually-annotated 150 job posts.

4.2.2 OBIE Evaluation & Discussion. In this experiment, OBIE was executed on the original texts for the same sample of 150 posts, containing a total of 1,760 sentences and 53,577 tokens. The annotation set generated for the OBIE pipeline was compared to that generated by the humans annotators. The Corpus Quality Assurance tool in GATE was used to calculate precision, recall, and the F_1-score between the two annotation sets. We report both the strict and lenient F-measure for the comparison.

On average, the F-measure ranges between 79-83%. This is satisfactorily close to the IAA rate of 94%. Also, both F1-strict and F1-lenient results are well within the reasonable performance range explained in related literature for IE tasks of a similar complexity [2]. The full results, summarised in Table 3, show that the lowest F-score (59% strict, 60% lenient) is achieved for the `SkillTool` annotation. For the 86 total extracted annotations in this category, 37 are consistent in both sets, 16 are unique to the IAA set and 33 are uniquely identified by the OBIE method. The second lowest F-score (68% strict, 73% lenient) is observed for `SkillProduct`. For the extracted annotations in this category, 153 are consistent in both sets (including 12 overlapping annotations), 93 are unique to the IAA set and 16 are uniquely identified by the OBIE method.

An investigation of the results reveals the following four major possibilities for improving this result:

- The standard GATE/Annie service occasionally fails to annotate skills,
- Ambiguity remains a challenge, e.g., *Hudson* in *"Hudson Global Resources Limited offers the services of an employment agency ..."*, *excel* in *"If this role sounds like something which you could excel in, please do not hesitate ..."*.,
- Skills unknown to the SARO instances are not annotated,
- Missing synonyms result in incomplete annotations, e.g., only *SQL* is marked up in *Microsoft SQL Server*, *MsOffice* and *MsAccess* are marked up, but *Microsoft Office* and *Access* are missed.

To address the above, we will observe a number of identified named entities that are not matched to skills, so as to continue populating the SARO ontology with new skill instances. We also reconsidered the JAPE Grammars to be more flexible with matching the text of named entities with skills (including lower and upper case letters). We also added general rules that add named entities with specific patterns as skills even when they are unknown to SARO, e.g., entities written in whole upper case (e.g.*CVS*, *GO*), starting in lower case and followed by digits (e.g.*k8s*), or consisting of a mix of upper case and lower case letters (e.g.*LeSS*, *SaaS*).

5 CO-WORD ANALYSIS AND DISCUSSION

The results of the OBIE method can be used as a basis for exploratory data analysis, different kinds of which are already available in the *European Data Science Academy* (EDSA) dashboard[11]. In this section, we investigate the viability of using these results to derive additional insights which are not yet provided, particularly time series analysis. Our target beneficiaries remain, primarily, i) applicants seeking to improve their competencies and personal competitiveness as data scientists, and ii) help educators fine-tune and update their curricula to meet demand from the hiring sector. We provide a proof-of-concept by way of a few examples derived from our sample dataset, and discuss the potential of the extracted data for identifying generalized skill demand composition and trends.

A data preparation stage is necessary to prepare the results for more advanced features over a timed series of large observations. To abstract the OBIE results, we consider co-word analysis to reduce data at any point into more workable clusters. In particular, in section 2 we identified the Paris/Keele method (subsection 3.3) as the most appropriate means for restricting the cluster size, by varying the permitted values for co-occurrence and the number of permitted links.

After considering the results of co-word analysis, we identified some generic skills which contribute little to the analysis while skewing the results considerably. For this study, we ignored four

[11] http://edsa-project.eu/resources/dashboard/

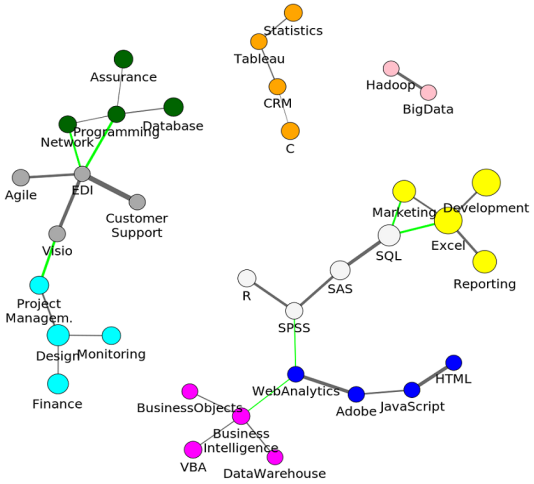

Figure 3: Skills' network with 1st variant

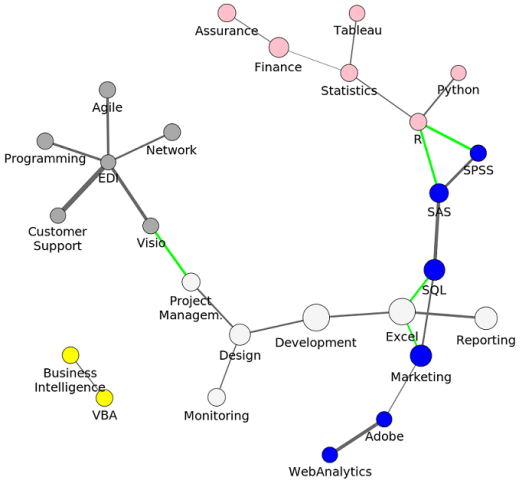

Figure 4: Skills' network with 2nd variant

frequently-occurring field words (*Analyst*, *Analysis*, *Analytics*, *Analytic*) (see Table 1), leaving 180 skills under consideration.

The final co-occurrences observed ranged from 1 (1087 skill pairs were only observed once in the job post sample) to 167 (1 skill pair was observed in 167 job posts). In order to generate our examples, we selected two variations of the Paris/Keele method. The first was given a co-occurrence threshold ≥ 10, and Pass1Links and MaxLinks coefficients equal to 3 and 5, respectively. This threshold range was chosen to generate clusters that are not incomprehensibly large and also not bound with very weak links. It ignores skill pairs occurring in less than 10 job posts. In a second variant the threshold set to ≥ 15, and Pass1Links and MaxLinks of 5 and 8, respectively.

The resulting skill networks are shown in Figure 3 (first variant) and Figure 4 (second variant). The former shows nine sub-networks, whereas the latter consists of five. The gray lines connecting nodes represent Pass1Links and green lines represent Pass2Links. The relative size of nodes represents their frequency of occurrence, and the relative size of lines represents the association strength between keywords.

The two versions of the skill networks shown demonstrate how the most prevailing complementary skills observed in skills demand data can be identified. These clusters can guide the design of curricula, or the identification of career pathways for aspiring data scientists targeting specialized niches within the field.

In addition to the skill maps, strategic diagrams can also provide more specific insights. The two diagrams corresponding to the two variants above are shown in Figure 5 and 6. The clusters identified by the network are plotted into four quadrants, enabling further observations. Below, we describe some of these observations for variant 1 (Figure 5).

Quadrant I has the highest centrality and density which indicates that clusters 1 and 3 are considered crucial and an important sub-area of skills by the employers. High density shows that these clusters are internally coherent and that their skills tend to be mature and developed. High centrality indicates that these clusters are strongly connected to other clusters. That is, skills in clusters 1 and 3 are the core data science skills. In contrast, cluster 5 is also

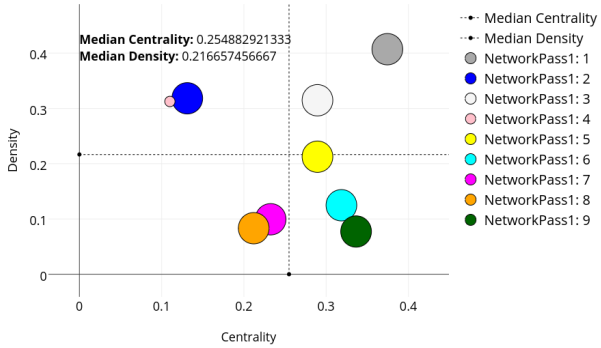

Figure 5: The strategic diagram with 1st variant

close to quadrant I, indicating that it has a higher tendency to be well-developed and become part of the core data science skills.

In quadrant II, cluster 5, 6, and 9 have high centrality and low density. That is to say, although the skills in these clusters are also core to the data scientist role, they require more investment and in-depth preparation. Note that more detailed information and knowledge, which allows us to determine the skill contribution to this field, is possible with a dynamic analysis (the evolution of a network over several periods, i.e., time series analysis) or a comparative one (the relationship of the network with other networks).

Quadrant III is characterized by low centrality but high density, hence clusters 2 and 4 are not central but have a close, internal connection. As with *quadrant II*, more insights can be extracted with the help of time-series analysis or network comparison.

Quadrant IV includes clusters 7 and 8, which reveal that the skills in these clusters are marginal and less developed, as stated by the low density and centrality. As with *quadrant II* and *III*, time-series study is also needed to get deeper insights into the contribution of these skills to the field.

The assessment of demand-derived technical skills provides our targeted users with useful direction because: (1) these findings could

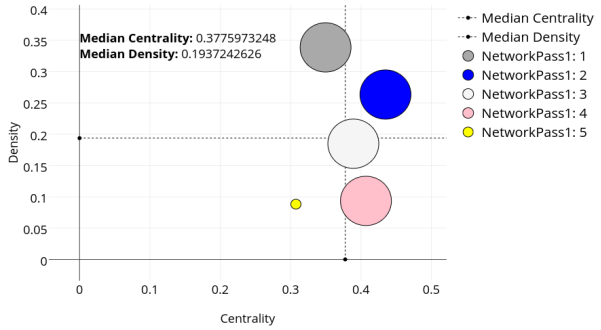

Figure 6: The strategic diagram with 2nd variant

aspire and reorient job seekers and applicants toward the pursuit of knowledge of demanded skills, to strengthen their existing skills or obtaining technical education; and (2) the implications of this new information could also motivate the educators and training providers to activate professional development, create a new course, or revise a course.

The constellation of the correlated skills presented above (e.g. Figure 3) and its global representation of the skill sub-networks (e.g. Figure 5), specifically in *quadrant I*, describes key areas of technical competencies currently needed in data science, from *EDI* and *Customer Support* to *SAS* and *SQL*. In addition, to receive more insight concerning skills located in *quadrant II, III,* and *IV*, and also to detect the difference among networks (Figure 5 and 6), a time-series study or a network comparison, planned as part of our future efforts, will help to trace shifts and changes in the labor demand.

6 CONCLUSION AND FUTURE WORK

In this article we presented an ontology-based information extraction method, guided by a domain vocabulary (SARO), that is able to identify data science skills in job posts. In our evaluation, the performance of the automatic extraction method is compared to its manual (human) equivalent, whose upper bound was measured at around 94% F-measure during the creation of a gold standard and the calculation of the human inter-annotator agreement. With 79%-83%, the resulting F-measure for the automated extraction fares very well and proves the feasibility of the method. The experiment also identified a number of limitations with the ontology and the method that can be addressed in future research to improve the performance.

In addition to the evaluation of the vocabulary-guided skill extraction, we included a study which also serves as a proof-of-concept validating the value and potential of the extracted demand data to identify skill demand composition and trends, so as to guide aspiring data scientists and educators alike. A sample of the collected data underwent a co-word analysis using the most suitable method identified, to generate abstracted skill maps based on co-occurrences, rather than raw frequency counts, which compensates for large differences in counts of commonly occurring terms [12]. In addition, the use of ontologies has added the benefit that both the raw elicited information, as well as higher-level abstractions and results, are all available in a standard format (RDF), thus enabling additional querying and analysis to be performed by third parties.

Our future work will focus on improving the presented information extraction process, and in part on maximizing the value of the extracted demand data. In particular, we will rely on a series of observations, such as the ones presented, to perform time series analysis and identify trends and shifts in demand skills by the relevant job sectors. The selection of an appropriate time series analysis method will be complemented by a Web-based User Interface offering insights into trends over custom periods.

ACKNOWLEDGMENTS

Elisa Margareth Sibarani is supported by a scholarship of *Indonesia Endowment Fund for Education* (LPDP), and part of the work reported in this article was supported by the EU project EDSA (EC no. 643937).

REFERENCES

[1] Michel. Callon, Jean-Pierre. Courtial, and F. Laville. 1991. Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. *Scientometrics* 22, 1 (1991), 155–205.
[2] Hamish Cunningham. 2005. Information extraction, automatic. *Encyclopedia of language and linguistics*, (2005), 665–677.
[3] Hamish Cunningham, Diana Maynard, and Kalina Bontcheva. 2011. *Text Processing with GATE (Version 8)*. Gateway Press CA.
[4] Johan De Smedt, Martin le Vrang, and Agis Papantoniou. 2015. ESCO: Towards a Semantic Web for the European Labor Market. In *Proceedings of the Workshop on Linked Data on the Web, LDOW 2015, co-located with the 24th International World Wide Web Conference (WWW 2015), Florence, Italy, May 19th, 2015.* http://ceur-ws.org/Vol-1409/paper-10.pdf
[5] Ying Ding, Gobinda G Chowdhury, and Schubert Foo. 2001. Bibliometric cartography of information retrieval research by using co-word analysis. *Information processing & management* 37, 6 (2001), 817–842.
[6] Ray Harper. 2012. The collection and analysis of job advertisements: A review of research methodology. *Library and Information Research* 36, 112 (2012), 29–54.
[7] Qin He. 1999. Knowledge discovery through co-word analysis. *Library Trends* 48, 1 (1999), 133–159.
[8] Chang-Ping Hu, Ji-Ming Hu, Sheng-Li Deng, and Yong Liu. 2013. A co-word analysis of library and information science in China. *Scientometrics* 97, 2 (2013), 369–382.
[9] Mary Anne Kennan, Patricia Willard, Dubravka Cecez-Kecmanovic, and Concepción S. Wilson. 2007. IS Early Career Job Advertisements: A Content Analysis. In *Pacific Asia Conference on Information Systems, PACIS 2007, Auckland, New Zealand, July 4-6, 2007.* 51. http://aisel.aisnet.org/pacis2007/51
[10] Mary Anne Kennan, Patricia Willard, Dubravka Cecez-Kecmanovic, and Concepción S. Wilson. 2009. A Content Analysis of Australian IS Early Career Job Advertisements. *Australasian J. of Inf. Systems* 15, 2 (2009). http://journal.acs.org.au/index.php/ajis/article/view/455
[11] Chuck Litecky, Andrew Aken, Altaf Ahmad, and H. James Nelson. 2010. Mining for Computing Jobs. *IEEE Software* 27, 1 (2010), 78–85.
[12] Linda Marion, Mary Anne Kennan, Patricia Willard, and Concepción S Wilson. 2005. A tale of two markets: employer expectations of information professionals in Australia and the United States of America. (2005).
[13] Sonia Sánchez-Cuadrado, Jorge Morato, Yorgos Andreadakis, and José A. Moreiro. 2010. A study of labour market information needs through employers' seeking behaviour. *Inf. Res.* 15, 4 (2010). http://informationr.net/ir/15-4/paper441.html
[14] Ritesh Shah and Suresh Jain. 2014. Ontology-based information extraction: An overview and a study of different approaches. *International journal of computer Applications* 87, 4 (2014).
[15] Sami Surakka. 2005. Analysis of Technical Skills in Job Advertisements Targeted at Software Developers. *Informatics in Education* 4, 1 (2005), 101–122. http://www.mii.lt/informatics_in_education/htm/INFE055.htm
[16] Peter A. Todd, James D. McKeen, and R. Brent Gallupe. 1995. The Evolution of IS Job Skills: A Content Analysis of IS Job Advertisements from 1970 to 1990. *MIS Quarterly* 19, 1 (1995), 1–27. http://www.jstor.org/stable/249709
[17] John Whittaker. 1989. Creativity and Conformity in Science: Titles, Keywords and Co-word Analysis. *Social Studies of Science* 19, 3 (1989), 473–496.
[18] Daya C. Wimalasuriya and Dejing Dou. 2010. Ontology-based Information Extraction: An Introduction and a Survey of Current Approaches. *J. Inf. Sci.* 36, 3 (June 2010), 306–323.
[19] Izabela A. Wowczko. 2015. Skills and Vacancy Analysis with Data Mining Techniques. *Informatics* 2, 4 (2015), 31. http://www.mdpi.com/2227-9709/2/4/31