Information Systems 65 (2017) 1–6

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/is

Data mining approach to monitoring the requirements of the job market: A case study

Ioannis Karakatsanis[a,*], Wala AlKhader[b], Frank MacCrory[c], Armin Alibasic[b], Mohammad Atif Omar[b], Zeyar Aung[a], Wei Lee Woon[a]

[a] Electrical Engineering and Computer Science, Masdar Institute of Science and Technology, Abu Dhabi, United Arab Emirates
[b] Engineering Systems and Management, Masdar Institute of Science and Technology, Abu Dhabi, United Arab Emirates
[c] MIT Sloan School of Management, Cambridge, MA, United States

ARTICLE INFO

Keywords:
Latent semantic indexing
Text-mining
Job market analysis
Web data extraction

ABSTRACT

In challenging economic times, the ability to monitor trends and shifts in the job market would be hugely valuable to job-seekers, employers, policy makers and investors. To analyze the job market, researchers are increasingly turning to data science and related techniques which are able to extract underlying patterns from large collections of data. One database which is of particular relevance in the presence context is O*NET, which is one of the most comprehensive publicly accessible databases of occupational requirements for skills, abilities and knowledge. However, by itself the information in O*NET is not enough to characterize the distribution of occupations required in a given market or region. In this paper, we suggest a data mining based approach for identifying the most in-demand occupations in the modern job market. To achieve this, a Latent Semantic Indexing (LSI) model was developed that is capable of matching job advertisement extracted from the Web with occupation description data in the O*NET database. The findings of this study demonstrate the general usefulness and applicability of the proposed method for highlighting job trends in different industries and geographical areas, identifying occupational clusters, studying the changes in jobs context over time and for various other research embodiments.

1. Introduction

1.1. Background

Rapid changes in working environments and newly adopted technologies have a profound impact on economic development [1]. While technological progress has generally augmented productivity, it has also resulted in structural unemployment, inequality and social imbalance [2]. In such uncertain economic setting standardized data about the skills and attributes of occupations is a valuable asset for firms, managers and employers to access basic information about the requirements and needs of the job market [3].

O*NET is a publicly available database that has been used by many researchers for examining the changes in occupational skill composition and identifying significant skill categories. This analysis has been proved very useful for understanding how the skills and abilities required for each job have been affected by recent technological advancements changes. Many of these studies have been centered around the changes in the intensive margin which focuses changes in the composition of jobs themselves [2,4,5]. Analysis of the extensive margin (i.e. changes in the demand and distribution of jobs) is also of great interest to the scientific community but the format of the O*NET database would be less appropriate in this case. This is because it was designed to characterize the skills required for different occupations but does not actually characterize the demand or prevalence of those jobs. For example, "Project Manager" and "Software Developer" would both appear as a single line in the O*NET database and be equally weighted, whereas companies will typically employ many more developers than project managers. A further limitation is that O*NET was developed for the US job market and might contain terms and job titles that are US specific. This in turn limits the applicability of O*NET when analyzing international job markets.

1.2. Motivation and objectives

National agencies publish structured data related to labor market demand, but these data lack sufficient detail or timeliness to suit our objective. For example, the US Bureau of Labor Statistics publishes the

* Corresponding author.
E-mail addresses: ikarakatsanis@masdar.ac.ae (I. Karakatsanis), ikarakatsanis@masdar.ac.ae (W.L. Woon).

http://dx.doi.org/10.1016/j.is.2016.10.009
Received 9 October 2016; Received in revised form 23 October 2016; Accepted 27 October 2016
Available online 29 October 2016
0306-4379/ © 2016 Elsevier Ltd. All rights reserved.

Occupation Employment Survey once per year and groups occupations at a coarser level than O*NET. Some nations of interest do not publish comparable statistics at all. This paper addresses detail, timeliness and US-specificity by suggesting a Latent Semantic Indexing (LSI) model for matching job postings existing on the Web with occupation data in the O*NET database. Specifically, we aim to detect the O*NET occupations with the greatest demand, identify popular job clusters and study the changes over time in the job market. A link to the O*NET data like this will allow direct comparisons to be made between various markets consisting of jobs with different names and roles in different countries. Such analysis can be of very great value to human resource managers, organizations, recruiting employers and job seekers trying to understand the skills requirements in the modern job market era. In addition, the increasing number of studies which specifically target the skills content of jobs [2,5] requires a specific O*NET database to be available for each country, region etc. Since this is a difficult task that has to be repeated for each country developing a method that maps different markets to O*NET could offer a viable alternative.

2. Data collection and pre-processing

2.1. O*NET

O*NET contains detailed information on more than 1000 US occupations. Among this information the fields that are most commonly used by previous studies are those which describe the skills, abilities and typical work activities corresponding to each occupation [2,5]. These fields are expressed as numerical values reflecting highly trained labor experts' assessments of the significance of each field to each occupation. However, the present work follows a different approach by utilizing textual information of fields describing the tasks, duties and requirements associated with each occupation. Hence, fields contain information like Occupation Title, Occupation Description and Tasks associated with each occupation were used from the O*NET to estimate the similarity between each specific job advertisement found online with the set of occupations existing in the database. It is important to note that O*NET only contains occupation types and not the distribution or actual frequency of these occupations. This is what we aim to quantify by matching each occupation in the database with job advertisements appearing at popular career websites.

2.2. Job postings

Job adverstisements were collected from a number of online jobs sites[1]. Since our primary goal is to study the demand of jobs in different industries existing at the same or different geographic region we collected 3814 job postings from the oil and gas industry in the GCC countries (United Arab Emirates, Saudi Arabia, Oman, Qatar, Bahrain, Kuwait), and 1423 unique job postings from the banking and finance sector in the same region. We also collected 3787 jobs' advertisements in the oil and gas industry in the following 5 US states: Texas, California, Louisiana, Oklahoma and Pennsylvania. According to American Petroleum Institute [6] these are the top 5 states in terms of the total number of jobs directly or indirectly attributable to the oil and gas industry. Also, the number of the jobs found here were very similar to the number of jobs from the GCC. As with the GCC case, we also extracted job postings for the banking and finance sector for these 5 US states. However, in this case the search yielded a much larger number of jobs than the 1423 found in the GCC area so a random subset of jobs was sampled from the total pool of jobs. To investigate further the generalisability of our method, we collected 1720 job postings from the oil and gas industry in UK (England, Scotland, Wales and Ireland), and 2105 distinct jobs' advertisements from the banking and finance market in the same area.

[1] Indeed.com, Monster.com, Naukrigulf.com, Jobsite.co.uk, Totaljobs.com. Collection period: July to November 2015 for GCC and USA, October to November 2016 for UK.

2.3. Pre-processing

Following common text-mining techniques, we removed common expressions and words from our data (both O*NET occupation descriptions and job advertisements) to reduce its size and noise. We tokenized and deleted terms appearing only once as well as we reduced the vocabulary produced by deleting stop words (e.g "and", "is", "thus"). The next step was to manually review the remaining corpus by filtering out irrelevant terms or keeping in terms which incorrectly were classified as invalid english words by the software used. After this manual data clean-up, the last step was to remove morphological affixes from words, leaving only the word stem, a process commonly known as stemming (i.e a stemming algorithm reduces the words "studying", "studies", and "student" to the root word, "study").

3. Proposed methodology

The next sections describe how we applied LSI to estimate the demand of O*NET occupations using job advertisement documents. Fig. 1 illustrates the procedure followed.

3.1. Latent semantic indexing (LSI)

Latent Semantic Indexing (LSI) is based on the idea that there are hidden semantic relationships in the corpus data and that the connection between words and documents can be better elucidated in this underlying space [7].

Initially, the O*NET occupation descriptions have to be transformed into vectors to form a vector space. Each element in a vector represents a word. In this way, the textual data can be represented by a word-document matrix X ($n \times m$), where n is the size of the terms used and m the size of O*NET occupation descriptions as it can be seen in Fig. 1. Each element x_{td} of matrix X represents the prevalence of term t in document d, weighted using the Term Frequency-Inverse Document Frequency (TF-IDF) formula. Given a document corpus D, a term t and a single document so that $d \in D$, we calculate weights as follows [8]:

$$t_d = f_{t,d} \times \log(|D| f_{t,D}) \quad (1)$$

where $f_{t,d}$ is the number of times the term t appears in document d, $|D|$ is the size of the collection, and $f_{t,D}$ equals the number of documents in which t appears in D.

The next step of a general LSI model is to perform a Singular Value Decomposition (SVD) on the matrix to identify the concepts/topics contained in the text (see Fig. 1). The SVD of our matrix X, is the product of three matrices:

$$X = LSR^T \quad (2)$$

where L and R are the matrices of the left and right singular vectors and S is the diagonal matrix of singular values. The diagonal elements of S are ordered by magnitude, and therefore these matrices can be simplified by setting the smallest k values in S to zero. The columns of L and R that are associated to the elements of S that were set to zero are removed. The new product of these three matrices is a matrix X_k that is an approximation of the term-by-document matrix. This new matrix represents the original relationships as a set of orthogonal factors (topics) where the term and document vectors can be seen as a "semantic space" [7]. The quality of the SVD depends on the number of factors that have to be used. To identify the optimal number of topics we manually tested several numbers of factors from 10 to 100 and qualitatively reviewed the results [9,10]. For each of these numbers we performed SVD to estimate term and document loadings for each topic. The number of topics k produced the best results was equal to 69. This looks like a quite reasonable number if we consider the relatively small

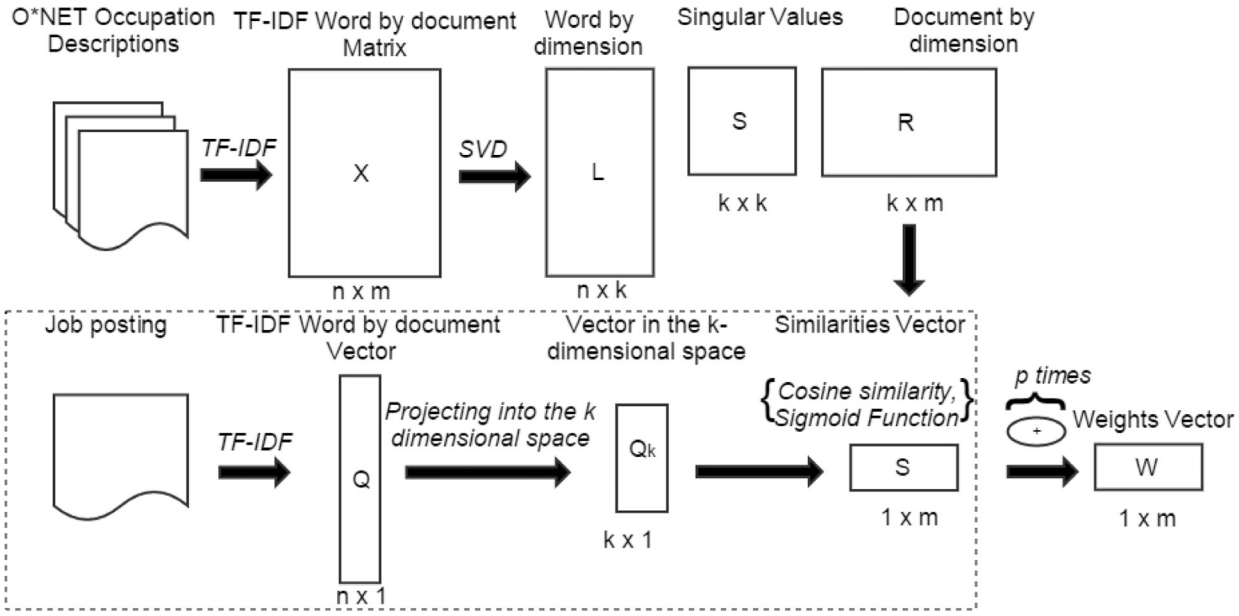

Fig. 1. Schematic diagram of the applied methodology. The number of similarity vectors produced is equal to the number of the job postings extracted. After normalizing each of these vectors we aggregate them into a single vector W containing demand weights for each O*NET occupation.

size of our trained data (1110 O*NET occupation descriptions).

3.2. Similarities and generation of weights

As explained earlier, LSI was used to transform the O*NET occupation data into a latent space of lower dimensionality. This will allow us to determine the similarity between a specific job posting extracted from the Web and each of the O*NET occupations. Each job posting is projected to a query vector Q_k in the same k-dimensional (k=69) semantic space that the O*NET occupation collection had previously been transformed into. This is achieved by using the following formula:

$$Q_k = S^{-1}L^T Q \tag{3}$$

where S^{-1} is the inverse of the diagonal matrix of singular values S, L^T the transposed of the left singular vectors matrix L and Q the term vector.

Once each job posting is represented in this way, the distance between a given query and occupation vector can be estimated using an appropriate distance measure. A number of different metrics were screened including the Pearson similarity, Euclidean distance, Bray-Curtis distance and Cosine similarity. After manually assessing a sample of the scores produced by all these metrics the Cosine similarity proved to be the most suitable. Given two vectors u_a and u_b their Cosine similarity can be computed as follows [11]:

$$CosSim(u_a, u_b) = \frac{u_a \cdot u_b}{|u_a| \times |u_b|} \tag{4}$$

u_a and u_b are n-dimensional vectors over the term set $U = u_1, \ldots, u_n$. Each dimension represents a word with its weight in the document, which is non-negative. As a result, the Cosine similarity is non-negative and bounded between [0,1].

However, the similarity scores produced this way cannot be used to efficiently highlight and detect occupational clusters. That's because for each occupation the corresponding Cosine similarity scores follow a very standard decline from values that are close to 1 and end up to 0. Thus, since our goal is to form a distribution of similarities for each occupation that clearly distinguishes the job advertisements that are similar to this occupation from those that are not similar we re-calculated each Cosine similarity using the Sigmoid function described in the following equation [12]:

$$Sigmoid(x) = \frac{L}{1 + e^{-k(x - x_0)}} \tag{5}$$

where e is Euler's number, x_0 the x-value of the sigmoid's midpoint, L the curve's maximum value and k the steepness of the curve. This approach resulted to scores between each job advertisement and the O*NET data that exhibit a progression from higher values close to 1 that accelerate and approach to 0 over time.

The outcome of the previously described procedure is a similarity vector S ($1 \times m$) where m is the number of O*NET occupations (1110 distinct occupations were used). As Fig. 1 shows, each time we generate p such vectors S where p is the number of the job postings extracted for the industry and geographic area our method is applied on (i.e for the banking and finance in GCC area n=1423).

To obtain a demand weight for each single occupation we first normalize each similarity vector S using Eq. (6) and then sum all p vectors into a single vector W. Each element of this vector is a demand weight for each occupation of the O*NET database.

$$\hat{s}_{1,j} = \frac{s_{1,j}}{\sum_{j=1}^{d} s_{1,j}} \tag{6}$$

3.3. Validation

Crowdsourcing was used to validate our proposed LSI model. For this purpose, CrowdFlower[2] as a service platform was deployed which allows users to access an online workforce of millions of people to clean, label and enrich data.

Initially, a random sample of the extracted job postings was selected in form of similarity vectors S. As described earlier, each vector contains similarity scores between the job posting being evaluated and the 1110 O*NET occupations. For each of those vectors two occupations were randomly chosen: one with corresponding similarity score greater or equal to 0.9 and another one with score lower than 0.9. Then, using CrowdFlower platform we created a task where people had to decide for each job posting which of the two given O*NET occupations (for obvious reasons their similarity scores were hidden)

[2] crowdflower.com

was the most similar to it. According to [13] higher precision can be achieved when humans have to decide among two options than when a plethora of selections is provided.

To ensure the quality of the answers, 39 test questions were created and every participant that did not satisfy a threshold automatically generated by the system was dismissed from the task. The experiment was running for approximately two hours and 750 trusted responses in total were obtained. A final report that aggregated all the responses was downloaded showing that 95.5% of the participants selected the O*NET occupation for which our LSI model assigned similarity score ≥0.9 as the most similar to the given job posting. These findings demonstrate that the proposed LSI model succeeds at identifying the strong correlations existing between the jobs advertisement documents and the corresponding O*NET occupation data.

3.4. Implementation details

All the analysis in this paper was conducted in Python. A number of libraries and toolkits were used where specialized functions were required:

1. The BeautifulSoup library [14] was used for collecting and parsing the job data from the Web.
2. NLTK (Natural Language Toolkit) [15] for preprocessing the corpus.
3. For LSI, we used the implementation provided in Gensim [16], which also offers a variety of built-in functions for text mining purposes.

4. Results and discussion

Due to space constraints, Tables 1, 2, 3, 4, 5, 6 show only the top 10 prevailing O*NET occupations identified by the proposed method for the oil and gas industry in GCC countries, USA and UK as well as for the banking and finance sector in the same three regions.

As was described in Section 3.2, the degree of "match" returned by the LSI model was used to infer the appropriate weighting for a given occupation. So for example, in Table 1 for the case of GCC countries, the occupation "Civil Engineers" has a higher weighting compared to that of "Mechanical Engineering Technologists" from which it may be inferred that more Civil Engineering than Mechanical Engineering positions are available in the oil and gas industry, and hence that the corresponding skills would be in greater demand. Note that this line of analysis is not concerned with the economic value of any of these jobs, merely the extent to which the skills related to these jobs are in demand in the market.

Another component in this line of analysis is the generation of intuitive visualizations using the generated weightings. We tried a number of methods but found that reasonable results could be obtained by applying Principal Components Analysis to the matrix of occupations and skills, and then overlaying the above-mentioned weightings to generate heatmaps. Figs. 2, 3, 4, 5, 6 and 7 show heatmaps for the different industries (oil and gas vs. banking and finance) and geographical areas (GCC countries USA, UK) that our method was applied on. Each heatmap is visualized using the first two principal components from the occupations-skills matrix. This matrix was constructed using information in the O*NET database by combining the importance levels for three main job characteristics for each occupation: *Activities, Abilities and Skills* [2,5].

Areas which are in high demand (large weights) are marked with a

Table 1
Top 10 most demanded O*NET Occupations in Oil & Gas industry in GCC countries.

Rank	Demand Weight	Occupation Title
1	23.86	Architectural and Engineering Managers
2	22.84	Mechanical Engineers
3	22.30	Administrative Services Managers
4	21.92	Civil Engineers
5	21.43	Electrical Engineers
6	21.27	Electronics Engineers, Except Computer
7	21.07	Petroleum Engineers
8	20.76	Aerospace Engineers
9	20.29	Mechanical Engineering Technologists
10	20.16	Civil Engineering Technicians

Table 2
Top 10 most demanded O*NET Occupations in Oil & Gas industry in USA.

Rank	Demand Weight	Occupation Title
1	25.30	General and Operations Managers
2	24.64	Transportation Managers
3	24.27	Administrative Services Managers
4	22.73	Chief Executives
5	21.81	Architectural and Engineering Managers
6	21.77	Management Analysts
7	20.99	Industrial-Organizational Psychologists
8	20.39	Mechanical Engineers
9	20.22	Sales Engineers
10	20.20	Human Factors Engineers and Ergonomists

Table 3
Top 10 most demanded O*NET Occupations in Oil & Gas industry in UK.

Rank	Demand Weight	Occupation Title
1	18.61	Search Marketing Strategists
2	16.96	General and Operations Managers
3	16.62	Sales Engineers
4	16.23	Industrial-Organizational Psychologists
5	15.74	Online Merchants
6	14.84	Sales Representatives, Wholesale and Manufacturing
7	14.72	Electronics Engineers, Except Computer
8	14.70	Architectural and Engineering Managers
9	14.69	Chief Executives
10	14.30	Marketing Managers

Table 4
Top 10 most demanded O*NET Occupations in Banking & Finance industry in GCC countries.

Rank	Demand Weight	Occupation Title
1	11.80	Sales Agents, Financial Services
2	11.49	Financial Managers, Branch or Department
3	11.44	Sales Agents, Securities and Commodities
4	11.20	Chief Executives
5	10.96	Financial Examiners
6	10.81	Spa managers
7	10.76	Investment Underwriters
8	10.75	Accountants
9	10.73	Actuaries
10	10.67	Financial Quantitative Analysts

Table 5
Top 10 most demanded O*NET Occupations in Banking & Finance industry in USA.

Rank	Demand Weight	Occupation Title
1	13.52	Sales Agents, Financial Services
2	12.37	Brokerage Clerks
3	12.03	Spa managers
4	11.84	Financial Managers, Branch or Department
5	11.79	Chief Executives
6	11.60	Sales Agents, Securities and Commodities
7	11.15	Actuaries
8	11.11	General and Operations Managers
9	11.03	New Accounts Clerks
10	10.68	Financial Examiners

Table 6
Top 10 most demanded O*NET Occupations in Banking & Finance industry in UK.

Rank	Demand Weight	Occupation Title
1	25.59	Sales Agents, Financial Services
2	24.28	Sales Agents, Securities and Commodities
3	24.19	Financial Managers, Branch or Department
4	22.72	Credit Counselors
5	21.67	Financial Examiners
6	21.66	Investment Underwriters
7	21.60	Accountants
8	21.43	Spa Managers
9	21.07	Treasurers and Controllers
10	20.99	Brokerage Clerks

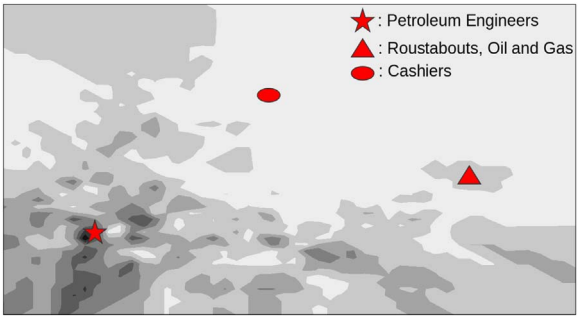

Fig. 2. The landscape of oil and gas industry jobs in GCC region.

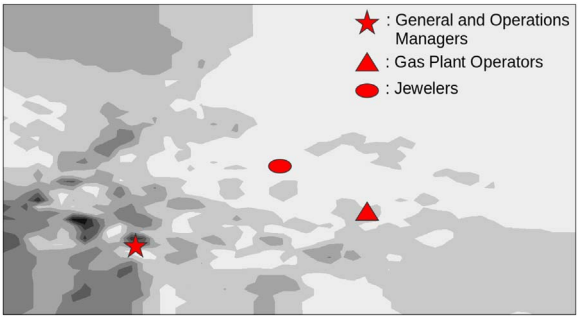

Fig. 3. The landscape of oil and gas industry jobs in USA.

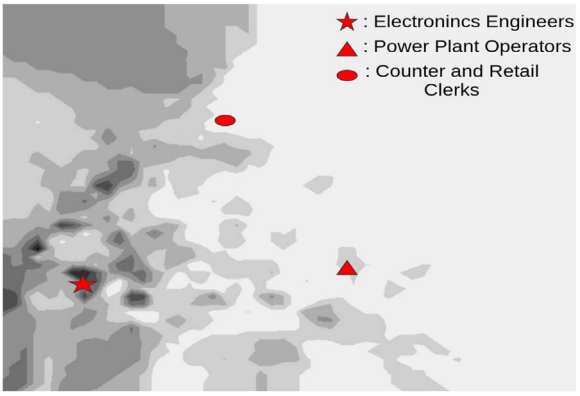

Fig. 4. The landscape of oil and gas industry jobs in UK.

dark grey color, low demand areas in light grey while intermediate levels of demand are assigned varying shades of grey. The resulting heatmaps contain all of the occupations in O*NET and are too big to fit into the current format. Hence, in each heat-map we only depict three occupation titles where each one belongs to a different group of weights and is representative of the results obtained for each case. Occupations belonging to high demand areas are coded with a “star” symbol,

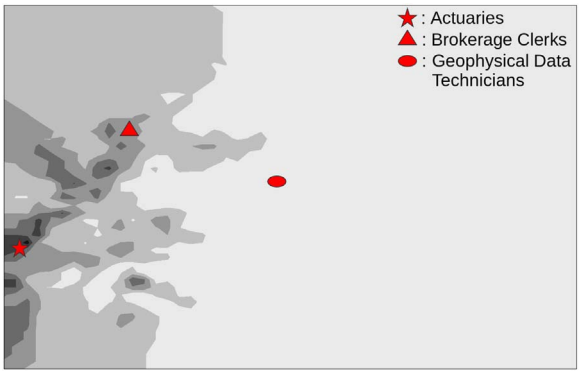

Fig. 5. The landscape of banking and finance industry jobs in GCC region.

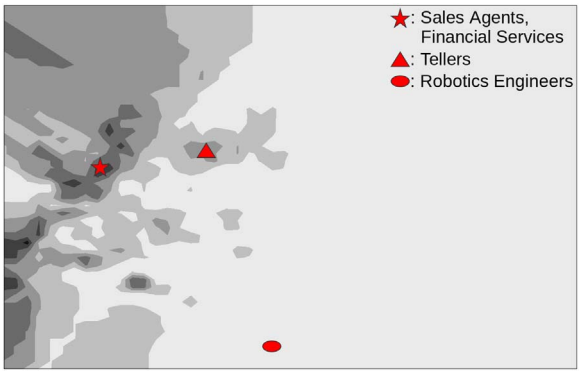

Fig. 6. The landscape of banking and finance industry jobs in USA.

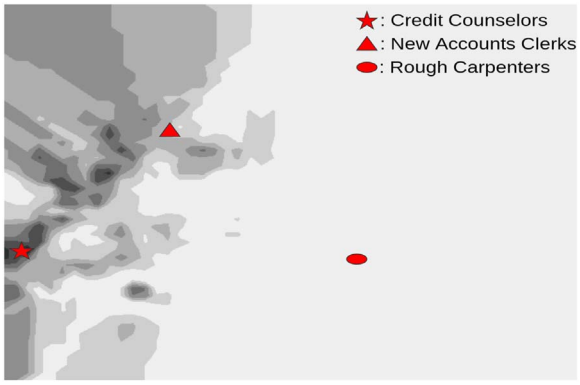

Fig. 7. The landscape of banking and finance industry jobs in UK.

occupations in the medium demand group with a “triangle” one while occupations in low demand are denoted with a “circle”.

Based on the results some observations are:

1. In comparison, the ten most in demand occupations in the oil and gas and in the banking sectors are clearly different. In Tables 1, 2, 3, many engineering professions are heavily weighted whereas this is noticeably less so in the banking sector, as can be seen from Tables 4, 5 and 6.
2. The occupation heatmaps strongly support the previous observation: while there is clearly a large difference between heat maps of the two different industries, heatmaps of the same industry but of different countries are broadly similar, though there are also some differences.
3. From Tables 2 and 3 there appears to be a greater managerial and/or administrative role for US and UK branches of oil and gas companies, while recruitment for the same companies operating in the gulf region (see Table 1) seem to focus more on technical and

hardware aspects, as may be expected.

4. As stated above, the differences between Figs. 2, 3 and 4 are much less pronounced but they are also not identical. This makes it harder to pick out broad differences between these figures but it does look like Figs. 3 and 4 show a slightly greater diversity of occupations (for e.g. the upper left region has greater weight in both Figs. 3, 4). This reflects a greater diversity of roles recruited for within the US and UK market, which would be expected as operations here cover a broader portion of the value chain ranging from exploration and production to managerial, finance and services. Additionally, the bottom right region of Fig. 4 is less weighted compare to that of Figs. 2 and 3 supporting the empirical assumption that the demand for technical and engineering roles in UK within the oil and gas sector is low.
5. The banking and finance market for GCC countries is very similar to that of USA and UK since the columns of Tables 4, 5 and 6 as well as the corresponding heatmaps shown in Figs. 5, 6, 7 look almost the same.
6. However, one interesting observation is the presence of the occupation "Spa managers" in Tables 4, 5 and 6. Initially, this may seem a little strange, but if we were to look up the corresponding entries in the O*NET, it will be seen that the database includes a list of related occupations (i.e "Spa Managers" are in a group, which lists the group of "General and Operations Managers" as related) that share the same skills. In addition, the skills corresponding to "Spa Managers" are quite generic (the top five are *Speaking, Monitoring, Coordination, Management of Personnel Resources, Service Orientation*), and indeed are likely to be shared by many jobs in the both oil and gas, and banking sectors. This illustrates how this form of analysis targets *skills* that are in demand, and not just specific occupations.

5. Conclusion and future work

In the current unpredictable economic environment occupational data about the skills and attributes provided by databases like O*NET is an essential source of information for managers, consultants and researchers for tracking changes in the job market over time. Although O*NET database contains a detailed skill set for a representative sample of occupations existing in the market, this data is not enough to characterize the distribution of the occupational skill demand. This paper proposed a Latent Semantic Indexing (LSI) model that utilizes O*NET occupational descriptions and raw job postings for identifying the most demanded occupations in the job market regardless the industrial sector and geographical area is focused on. Our findings reveal that this technique produces much finer and more timely readings of the labor market compare to the common official employment or job opening statistics. In addition, one really valuable property of the proposed method is that it can be directly applied to different job markets, commercial sectors and geographical regions - all that is needed is a corpus of jobs advertisements collected from the target market. In this preliminary study this was performed using web scraping scripts, but if more reliable or detailed results are required, customized data collection exercises can be performed by analysts.

The validation method presented here demonstrates the general applicability of the proposed LSI model. As future work, we aim to use our proposed data mining approach to conduct more in depth studies on the specifics of skills changes in jobs markets, and hence produce specific policy or strategy recommendations. In addition, further work might focus on improving the pre-processing methods in order to reduce the noise in the data as well as exploring more sophisticated approaches for normalizing and aggregating the generated occupation weights.

References

[1] A. Markusen, Targeting occupations in regional and community economic development, J. Am. Plan. Assoc. 70 (3) (2004) 253–268.
[2] F. MacCrory, G. Westerman, Y. Alhammadi, E. Brynjolfsson, Racing with and against the machine: Changes in occupational skill composition in an era of rapid technological advance.
[3] M.L. Hilton, N.T. Tippins, et al., A Database for a Changing Economy:: Review of the Occupational Information Network (O*NET), National Academies Press, 2010.
[4] H. David, F. Levy, R.J. Murnane, The skill content of recent technological change: An empirical exploration, Tech. rep., National Bureau of Economic Research (2001).
[5] W.L. Woon, Z. Aung, W. AlKhader, D. Svetinovic, M.A. Omar, Changes in occupational skills-a case study using non-negative matrix factorization, in: Neural Information Processing, Springer, 2015, pp. 627–634.
[6] API, Oil and natural gas stimulate american economic and job growth,(accessed 01. 11.15). URL ⟨http://www.api.org/~/media/files/policy/jobs/oil-gas-stimulate-jobs-economic-growth/api-state-vendor-survey-findings-report.pdf⟩
[7] S. Zelikovitz, H. Hirsh, Using lsi for text classification in the presence of background text, in: Proceedings of the tenth international conference on Information and knowledge management, ACM, 2001, pp. 113–118.
[8] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Inf. Process. Manag. 24 (5) (1988) 513–523.
[9] S. Debortoli, O. Müller, J. vom Brocke, Comparing business intelligence and big data skills, Bus. Inf. Syst. Eng. 6 (5) (2014) 289–300.
[10] N. Evangelopoulos, L. Visinescu, Text-mining the voice of the people, Commun. ACM 55 (2) (2012) 62–69.
[11] A.Huang, Similarity measures for text document clustering, in: Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand, 2008, pp. 49–56.
[12] P.-F. Verhulst, Notice sur la loi que la population suit dans son accroissement. Correspondance mathématique et physique publiée par a, Quetelet, 10, 1838, pp. 113–121.
[13] B. Schwartz, The paradox of choice: Why less is more, New York: Ecco.
[14] L. Richardson, Beautifulsoup, (accessed 02.07.15). URL ⟨http://www.crummy.com/software/BeautifulSoup/⟩
[15] S. Bird, E. Klein, E. Loper, Natural Language Processing with Python, O'Reilly Media, 2009.
[16] R. Řehůřek, P. Sojka, Software Framework for Topic Modelling with Large Corpora, in: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, ELRA, Valletta, Malta, 2010, pp. 45–50, ⟨http://is.muni.cz/publication/884893/en⟩.