# A NOVEL MULTI-MODAL FRAMEWORK FOR MIGRANTS INTEGRATION BASED ON AI TOOLS AND DIGITAL COMPANION

[1]*David Martín-Gutierrez,* [1]*Gustavo Hernández-Peñaloza,* [2]*Theodoros Semertzidis,* [1]*Francisco Moreno,*
[2]Michalis Lazaridis, [1]Federico Álvarez and [2]Petros Daras

[1]{dmz,ghp,fmg,fag}@gatv.ssr.upm.es, Universidad Politécnica de Madrid,
Madrid, Spain
[2]{theosem,Michalis.Lazaridis,daras}@iti.gr, Centre for Research and Technology Hellas,
Information Technologies Institute Thessaloniki, Greece

## ABSTRACT

ICT have proven to provide significant aid for appropriate integration of migrants. These tools can support the inclusion by providing guidance, education opportunities, job seeking, culture immersion and facilitating access to primary services. In this paper, a complete framework for migrants (with special focus on refugees) guidance and inclusion is presented. This framework comprises a set of novel AI tools aimed at enabling mentioned services from diverse perspectives: a) users' profiling; b) skills matching c) recommendations; d) user profiling and e) digital companion. Consideration about data collection, data flow, architecture and interactions are provided.

***Index Terms—*** AI users' profiling, Deep Learning.

## 1. INTRODUCTION

Migrations' flow has a significant impact at all levels for our societies [1]. On the one side, the effects on people arriving in a new, and sometimes unknown, place where the conditions are different and not having even mechanisms for interaction [2]. On the other side, consequences on the destination place, where the systems have not the adequate tools for integration yield to a high rate of migrants exclusion and co-lateral side repercussions.

In fact, the European Union is experiencing large migrant and refugee flows, both internal and external, due to the economic environment and socio-economic instability in the greater region, with large populations near or within war zones. To effectively manage the constant flows of migrants, national and local authorities and communities need to efficiently address the challenges that emerge in the management and integration of people hosted in their regions.

Lack of background information (including their legal status), communication difficulties and lack of trust are the key roadblocks for efficient integration. Migrants' cultural background, education and language skills are also playing an important role in their ability to fit in their new society.

Therefore, there is a need for holistic solutions involving all stakeholders (authorities, Non-Governmental Organizations and related institutions offering services to migrants (i.e. volunteers) & migrant communities).

In this context, ICT tools can contribute to face with these challenges by providing services to guide and foster the fast integration of migrants, supporting them in the access to assistance, orienting them in meaningful tasks such as housing, job seeking and education.

In this paper, an approach to guide migrants by a set of AI toolbox is presented. This set of micro-services will use migrants information inserted in multiple modalities to profile users and allowing them to be provided with personalized services based on skills matching and defined user needs. Further, interaction via AI conversational-agents services and perception of migration are considered.

The remainder of this paper is organized as follows: section 2 describes the State-of-the-Art of techniques employed for services provision, whereas section 3 details the architecture and main components of the proposed system. 5 describes the AI components proposed. Finally, 6 outlines the conclusions and Future Work.

## 2. STATE-OF-THE-ART AND RELATED WORK

The overall approach presented is split into a set of modules and interrelated technologies that use the information collected by the user (categorical variables and general related data), project this information onto a mathematical representations to extract a set of representative features, that are applied to complex techniques for diverse tasks such as ranking, classification, clustering and prediction. These techniques can be classified into 2 main groups: supervised and unsupervised. The former group of techniques rely on the use of labeled data-sets to create complex models able to "learn a task", being classification the most typical application, whereas in unsupervised techniques most representative task to be clustering. Therefore, according to the topics ad-

**978-1-7281-1331-9/20/$31.00** 

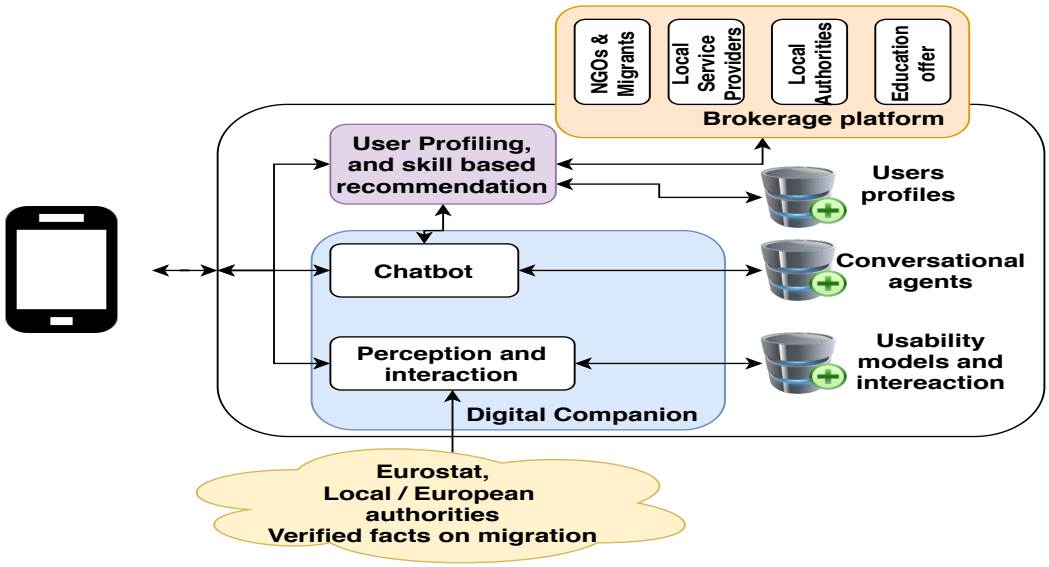


**Fig. 1**. General overview of the proposed architecture. The information collected by migrants will permit to create skill-matching based recommendations with related associations via brokerage platform, whereas the chatbot can support the migrant's guidance.

dressed, the state-of-the-art is reported as follows:

**Mathematical embeddings**: There is an broad literature to create unsupervised embeddings. One of the most known technique is fundamental for the Natural Language Processing techniques and is known as word embeddings (word2vec) was [3], which was the foundations of diverse subsequent *2vec algorithms. Moreover, several techniques have relied on graph embeddings, including classical approaches such as multidimensional scaling (MDS) [4], Laplacian, Eigenmap [5] IsoMap [6] or and Large-Scale Information Network Embeddings approach [7], and recent approaches such as graph factorization [8] or DeepWalk [9]. However, recent approaches using Neural Networks NN are becoming more popular. NN, especially in the light of recent advances that have made them the state of the art for many statistical tasks including learning to rank [[8, 10]].

**Skill Matching** In [11], a skills matching system is implemented, describing an architecture to collect data and evaluating several matching approaches. These algorithms are evaluated by using historic data of a company for 49 job positions. In [12], a skill matching algorithm is employed to increase the fitting of a candidate search for a particular vacancy. Several rankings are provided with the ambition of support human selection assessment. Some other approaches calculate the candidate's suitability without considering the relative ordering of the initial result list. In [13], regression algorithms are applied in a job market context to rank candidates for a job by fusing hard skills extracted from CVs with signals from soft skills found using social media.

**Recommendations**:There is a wide array of approaches in literature regarding the skills matching based recommendation. A set of solutions relies on description logic [14], Machine Learning (ML) [15], or semantic ontology-based approaches [16]. There are hybrid proposals that fuse multiple approaches aiming at improving the performance. As an example, in [17] ML algorithms on top of onthologies are presented.

## 3. SYSTEM ARCHITECTURE AND DATA-FLOW

From a high level perspective, the architecture is depicted in Figure 1, and it is composed by three main systems: *a)* a set of smart interfaces for the interaction with migrants, *b)* a brokerage platform to connect all stakeholders towards a holistic integration and *c)* an AI-toolbox that provides the mentioned services. This keystone system is composed by two subsystems: *a)* A Recommender system Pipeline and *b)* A Digital companion including a AI-based chatbot and a Perception and interaction module. The former provides useful recommendations based on critical needs of migrants (such as job seeking and skills matching), whereas the Digital companion guides the migrants into the services available. In the next section, these components is detailed.

## 4. RECOMMENDER SYSTEM PIPELINE

More specifically, Figure 2 shows the overall data workflow of the recommender system, where three main stages are performed:

1. A **Data Collection**, where migrants are asked for personal, educational and professional information consid-

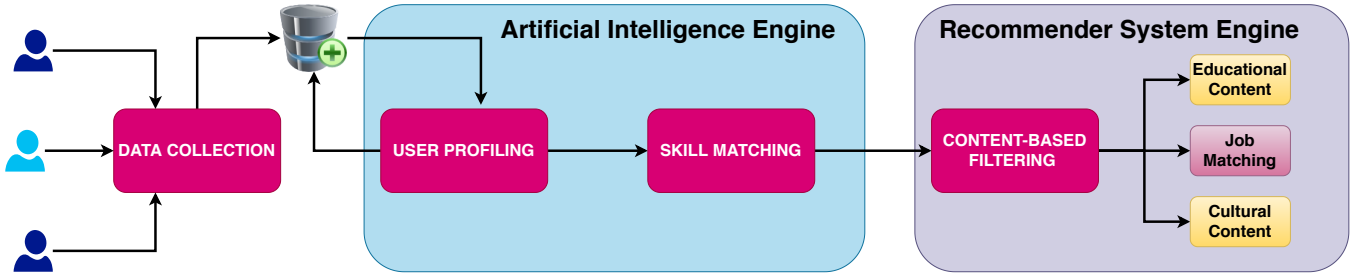


**Fig. 2**. An illustrative High-level block diagram showing the pipeline of the recommender system that attempts to facilitate both personal and professional life of migrants when they arrive to a new country using their previous professional experiences, skills and interests.

ering the General Data Protection Regulation (GDPR from now on) to follow the legal regulation.

2. An **Artificial Intelligence Engine (AIE)** which is composed by two main processes including a profiling and a skill matching.

3. A **Recommender System Engine (RSE)** which is responsible for sending recommendations to migrants based on the information provided by the AIE component.

### 4.1. Data collection

This component is responsible for retrieving a set of information from the different migrants via the global platform.

Firstly, migrants select their mother tongue and they are asked to fill out a very simple form to gather some specific and minimal information from them which can lead the system to automatically profile them and provide them with meaningful recommendations. Thus, the objective of this component is to maximize the performance of the system by minimizing the information that is needed from migrants in order to accomplish the terms established by the GDPR law. Respecting the GDPR restrictions, the proposed data gathering form has only optional fields except the username and an email for communicating information. However, if minimum information is provided, no recommendations are available but only collections of existing resources. More specifically, the following information is stored and modelled in the corresponding user database: i) a unique username ii) an email account iii) a range of age iv) the gender v) the mother tongue vi) additional spoken languages vii) the level of studies viii) the civil status ix) a previous Professional Job x) a list of professional skills including a range of expertise for each of them and finally, xi) a list of interests and hobbies.

Moreover, the aforementioned information is totally hashed and encrypted before being stored in the database to assure the data protection terms to final users. By applying this criteria, we are therefore attempting to encourage migrants and other users to employ the proposed system.

Furthermore, this process is totally dynamic, in the sense that, every time a new user logins the system, the data collection component is requested.

### 4.2. Artificial Intelligence Engine (AIE)

#### *4.2.1. Users Profiling*

The user profiling component attempts to perform two different goals:
**1.** Organise the set of users in clusters based on their profiles
**2.** Generate a low dimensional vector that better represents the information of a given user.

Thus, the system attempts to represent a given user $\Omega_i$ as a vector representation, so that $\Omega_i = [x_1, x_2, \dots x_N] \; \forall i = 0, \dots, M$, being $N$ the number of features and $M$ denoting the total number of users (migrants) in the database.

To address such purpose, the system performs an unsupervised learning procedure to classify the different users based on the similarities among them. More specifically, a considerable advantage of employing unsupervised learning is the absence of annotated data, which is both time-consuming and expensive.

One approach is connectivity-based algorithms known as hierarchical as [18, 19] describe. These algorithms produce an extensive hierarchy of clusters instead of a partition of the data points. They form clusters by connecting objects based on their distance and as the algorithm progresses the already created clusters are merged with each other. Depending on the strategy they use for the merging of the clusters they can be separated in *agglomerative* (bottom-up) and *divisive* (top-down) type.

Apart from the distance function that has to be defined, a linkage criterion is also necessary and the most common choices are single-link, complete-link, and average-link distance. A different approach that also uses a linkage criterion is based on density algorithms in which clusters are formed in areas with higher density than others.

Data points in the sparse area are considered as noise or border points. The most popular density based algorithm for

clustering is *DBSCAN* which is widely used in many applications as the ones described in [20, 21, 22].

Consequently, after performing several experiments, *DBSCAN* was selected to perform the user profiling procedure. We also apply the so-called *Silhouette coefficient* procedure as it is mentioned in [23].

As a result, after performing the clustering stage, the set of user profiles are mapped into a vector representation denoted as *embedding* which are finally stored in the database to be used for the rest of the components of the system.

#### 4.2.2. *Skills matching*

As it is mentioned in several studies such as [24, 25], the goal of a skill matching system consists in established a scoring metric which indicates the level or degree of matching between a given candidate and a job. Consequently, recruitment companies may retrieve the best candidates based on such scoring. However, in the proposed scenario the problem differ slightly from the aforementioned general approach in the sense that the objective of the problem consists in providing migrants and other users with the best set of jobs based on their profiles, whereas the original problem is the retrieval of the best set of candidates for a given job.

Thus, our approach follows an Ontology-based approach as the one described at [24], where a given user $\Omega_i \; \forall i = 1, \ldots, i = M$ is represented as a node in a graph of skills where all the relationships represent the set of skills associated to $\Omega_i$. We denote such set of skills as $S_\Omega = \{s_1, s_2, \ldots, s_K\}$.

Moreover, a set of weights $W_{(\Omega,S_\Omega)} = \{w_1, w_2, \ldots, w_K\}$ are employed to measure the importance of each skill. In this case, this importance value indicates the expertise that a given user $\Omega_i$ has for each $s_k$.

Additionally, a set of jobs $\Theta = \{\theta_1, \theta_2, \ldots, \theta_P\}$ is also involved in the graph representation, where each particular $\theta_p$ are nodes as well with a set of required skills $S_\Theta$ with their corresponding set of weights $W_{(\Theta,S_\Theta)}$ associated to them as well. All the weights are normalized to sum up to 1 in order to provide a final matching score between 0 and 1.

Hence, the algorithm attempts to search and match the subset of jobs that shares the highest number of skills with the given user considering also the weights to measure the level or degree of matching.

Figure 3 represents the skill-based graph that is employed in our proposed system to determine the best jobs that might be interesting for a given user according to the mutual connections that they have regarding the set of available skills $S$.

Furthermore, this approach is totally dynamic, in the sense that every time a new user profile is stored in the database, it is mapped into the graph, and the matching scores for each job are computed.

Subsequently, the algorithm sorts the matching score of all the jobs and retrieves a subset of $\Theta$ which represents the best $\beta$ jobs that have an adequate level of matching with the given user.

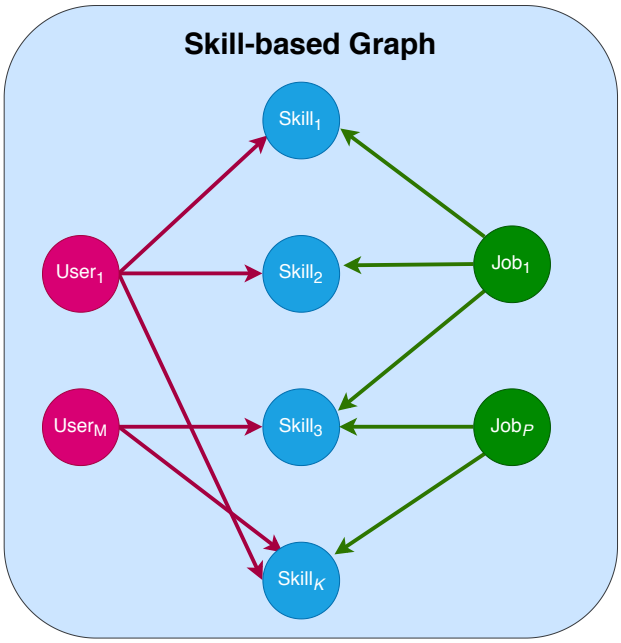


**Fig. 3**. Visual representation of the skill-based graph where the set of users $\Omega$, skills $S$ and jobs $\Theta$ are presented as red, blue and green circles respectively. On the other hand, the set of weights $W_{(\Omega,S_\Omega)}$ are showed as red lines whereas the set of weights $W_{(\Theta,S_\Theta)}$ are represented with green lines.

As expected, the output of this component are both the $\beta$ jobs and their corresponding matching scores. Finally, the output is sent directly to the RSE which will provide the pertinent recommendations to the given user.

### 4.3. Recommender System Engine (RSE)

In Section 4.2, a description of the procedures associated directly to the user profile are presented. This information is passed directly to the RSE component as main input in order to determine the recommendations that a given user must received. In particular, the RSE provides recommendations regarding three different topics including: **job matching**, **cultural content** and **educational content**.

#### 4.3.1. *Content-based filtering*

The input regarding the job matching recommendations is directly provided by the skill-matching component as we mentioned in Section 4.2.2. However, the remaining inputs are computed using both the user profile and the content that the proposed platform contains.

More specifically, to determine the best recommendations related to cultural content, the system considers the set of interests $I = \{i_1, i_2, \ldots i_K\}$ that the migrant or user has, and

the set of cultural content provided by the platform, which is denoted as $C = \{c_i, c_2, \ldots, c_L\}$. All this cultural content has to be properly annotated with the set $I$ of tags. Then, an interest-graph representation is determined to relate the connections between users and cultural content based on both the interests provided during the user profile stage and the tags of the cultural content.

The procedure for generating recommendations regarding educational content is very similar to the aforementioned method for generating cultural content recommendations. As before, all the educational content of the platform have been annotated beforehand with specific tags and skills that maps both the interests and skills which can be selected by users during the data gathering process. Therefore, in this scenario a skill-interest graph is generated to perform the analysis.

Thus, the recommendations provided by the RSE are content-based since they rely on the interests of the user according to their profiles. As expected, similar users will lead to have similar recommendations as a result of the proposed approach.

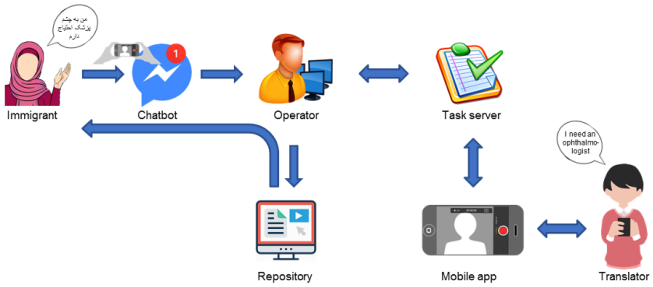


**Fig. 4**. The translation workflow of the video chatbot for the illiterate or not-supported languages

## 5. ADDITIONAL TOOLS FOR MIGRATION

### 5.1. Hybrid visual chatbot as digital companion

Communicating information in the right way is key challenge in the context of migration since the migrants population is extremely diverse in culture, language and educational level. The differences in culture and the vast number of dialects and languages migrants use makes the sharing of information extremely difficult. To make things more complicated, a large number of migrants are illiterate or their mother tongue has no written form.

Having all these restrictions as requirements, a digital companion in the form of a chatbot is proposed to drive the usage of the system by migrants. The basic conversations of the chatbot are using a rule-based conversational agent that is following predetermined conversation trees for specific information types such as book a medical appointment, get basic legal advice or move around the city, to name a few.

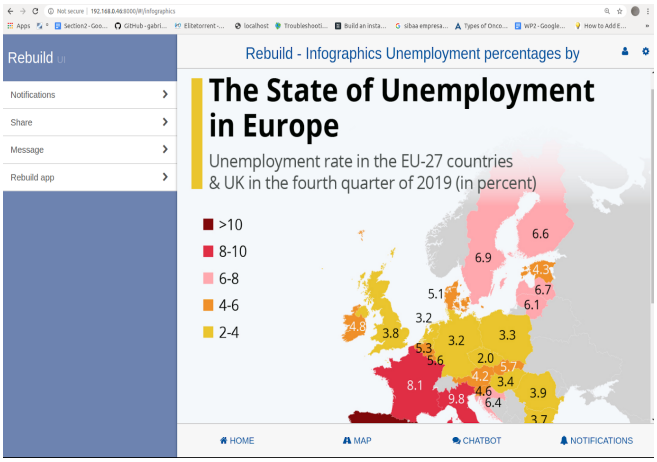


**Fig. 5**. A semantic search engine will permit to search from a set of reliable sources information related to the topics of interest to be displayed to improve real migration issues and unbiased perception

The chatbot is built on visual conversation trees i.e. conversation trees that are built around visuals and short videos. The information is enriched with text only for users that are able to communicate in written form. Moreover, a crowd based video exchange backend permits the asynchronous exchange of videos for migrants that are illiterate. Social workers or integrated migrants reply to these video messages to support the requests asynchronously. An example of such workflow is presented in Figure 4.

### 5.2. Migration perception and interaction

One of findings in European countries about migration perception is that the negative impression is mostly formed on the basis of fake premises. Although there is a huge amount of information, sometimes it is hard to classify the migration news and unfortunately, it yields to wrong judgment while not contributing to integration processes. For this purpose, a perception module is proposed. This module relies on the collection of reliable sources, which are crawled, and stored locally. Then, a search engine has been developed to extract from the knowledge database information related to topics of interest, as depicted in Figure 5.

## 6. CONCLUSIONS AND FUTURE WORK

In this paper, a complete framework for the integration of migration was presented. First, the data collection, special restrictions concerning GDPR were outlined. This framework relies on the use of advanced Skill recommendation schemes to provide a wide range of services to all stakeholders involved in migrants and refugees integration. From the multiple approaches, NNs for mathematical embedding and

graph-based skill-recommendations engine was selected. It looks for similarities on the migrants profile with the services offered (i.e. health, education, job seeking, migrant communities) via a brokerage platform. It is also interacting with a digital companion that comprises a visual chatbot and a perception and interaction modules. The former, supports visual (textual and graphic) guidance and recommendations, whereas the perception module is intended to double check information allowing to have a unbiased manner of interpreting the information about migration. As future work, the implementation of an AI-based support to guide migrants/refugees in a graphical-oriented manner can increase the scope of this framework.

# Acknowledgements

This work was supported by the H2020 European Project REBUILD https://www.rebuildeurope.eu/ Grant no. 822215.

## 7. REFERENCES

[1] D. Baltruks and A. Lara Montero, "The impact of the refugee crisis on local public social services in europe," *European Social Network*, 2016.

[2] E. Poptcheva and A. Stuchlik, "Work and social welfare for asylum-seekers and refugees: Selected eu member states," *EU Parliamentary Research Service*, 2015.

[3] K. Chen G. S. Corrado T. Mikolov, I. Sutskever and J. Dean., "Distributed representations of words and phrases and their compositionality," *NIPS*, 2013.

[4] T. F. Cox and M. A. Cox, "Multidimensional scaling.," in *Handbook of Data Visualization. Springer Handbooks Comp.Statistics*. Springer, 2008.

[5] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," *NIPS*, 2012.

[6] V. De Silva J. B. Tenenbaum and J. C. Langford., ," in *A global geometric framework for nonlinear dimensionality reduction.* Science, 2000.

[7] J. Tang et al, "Line: Large-scale information network embedding," *WWW*, 2015.

[8] S. Narayanamurthy V. Josifovski A. Ahmed, N. Shervashidze and A. J. Smola., "Distributed large-scale natural graph factorization," *WWW*, 2013.

[9] R. Al-Rfou B. Perozzi and S. Skiena., "Deepwalk: Online learning of social representations," *KDD*, 2014.

[10] Y. Shan et al, "Deep crossing: Web-scale modeling without manually crafted combinatorial features," *CIKM*, 2016.

[11] H. Braun, "Applying learning-to-rank to human resourcing's job-candidate matching problem: A case study.," *Master's thesis, Radboud Universiteit*, 2017.

[12] M. Fang, "Learning to rank candidates for job offers using field relevance models," *Master's thesis, University of Groningen & Saarland University*, 2015.

[13] Sondhi-P. Zhai C. Karmaker Santu, S.K., "On application of learning to rank for e-commerce search," 2017.

[14] et al A. Cali, ," in *A description logic based approach for matching user profiles.* in In Proc. of the 8th Int. Conf. on Knowledge Based Intelligent Information & Engineering Systems (KES), 2004.

[15] Evanthia Faliagka, Kostas Ramantas, Athanasios Tsakalidis, and Giannis Tzimas, "Application of machine learning algorithms to an online recruitment system," in *Proc. International Conference on Internet and Web Applications and Services*. Citeseer, 2012.

[16] M. Fazel-Zarandi and M. S. Fox, "Semantic matchmaking for job recruitment: an ontology-based hybrid approach," *Proceedings of the 8th International Semantic Web Conference*, 2009.

[17] M. Faheem F. M. Hassan, I. Ghani and A. A. Hajji, "Ontology matching approaches for erecruitment," *International Journal of Computer Applications*, 2012.

[18] Vincent Cohen-Addad, Varun Kanade, Frederik Mallmann-Trenn, and Claire Mathieu, "Hierarchical clustering: Objective functions and algorithms," *Journal of the ACM (JACM)*, 2019.

[19] Moses Charikar, Vaggos Chatziafratis, and Rad Niazadeh, "Hierarchical clustering better than average-linkage," in *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2019.

[20] K Mahesh Kumar and A Rama Mohan Reddy, "A fast dbscan clustering algorithm by accelerating neighbor searching using groups method," *Pattern Recognition*, vol. 58, pp. 39–48, 2016.

[21] Michael Hahsler, Matthew Piekenbrock, and Derek Doran, "dbscan: Fast density-based clustering with r," *Journal of Statistical Software*, vol. 25, 2019.

[22] Daren Wang, Xinyang Lu, and Alessandro Rinaldo, "Dbscan: Optimal rates for density-based cluster estimation," *Journal of Machine Learning Research*, vol. 20, no. 170, pp. 1–50, 2019.

[23] Hong Bo Zhou and Jun Tao Gao, "Automatic method for determining cluster number based on silhouette coefficient," in *Advanced Materials Research*. Trans Tech Publ, 2014, vol. 951, pp. 227–230.

[24] Teodor Petrican, Ciprian Stan, Marcel Antal, Ioan Salomie, Tudor Cioara, and Ionut Anghel, "Ontology-based skill matching algorithms," in *2017 13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*. IEEE, 2017.

[25] Jorge Martinez-Gil, Alejandra Lorena Paoletti, and Mario Pichler, "A novel approach for learning how to automatically match job offers and candidate profiles," *Information Systems Frontiers*, pp. 1–10, 2019.