Skill Job Recommender

Literature Survey

1.1 Introduction

The recommender system is becoming part of every business. The business tries to increase its revenue by raising the user's interaction by recommending new items based on user preferences. We have witnessed the rise of Netflix in the entertainment domain, using their strategies to implement a recommender system into their existing ecosystem. But there has been a minimal study in the hiring field from the perspective of a job seeker. To start any research, it is quintessential to review relevant work in the domain and technology.

1.2 Recommender Systems

As discussed previously, RecSys are the system that analyses user preference history and caters them with different options of services related to the requirement. Recommender systems emerged as an independent research area in the mid-1990s(Ricci *et al.*, 2011). In recent years, the interest in recommender systems has dramatically increased. In the Recommendation algorithm, it classifies into four types: Content-based filtering, Collaborative filtering, Rule-based, and Hybrid approaches (Mobasher, 2007; Al-Otaibi and Ykhlef, 2012).

Collaborative Filtering (CF): Collaborative Filtering is a technique is based on the human ratings that are given to an item by a user and find similarity between different users who have given similar ratings to an items(Hu and Pu, 2011). The essential operation used here is the memory-based nearest neighbor approach to group users who have a similar interest. As the volume of data grows gradually, there will be high latency in generating recommendations Mobasher (2007); Herlocker *et al.* (1999). Collaborative filtering has an advantage over content-based filtering techniques, but due to the nature of the hiring process, a job cannot be rated by the user and will not be possible to create a similarity matrix.

Content-based filtering (CBF): These are the most subjective and descriptive based filtering. Content-based filtering can also be called as attribute-based recommender as it uses the explicitly defined property of an item. It is an approach to an information retrieval or machine learning problem. The assumption made in content-based filtering is that user prefers item with similar properties. Content-based filtering recommends items to the user whose properties are similar to the item which the user has previously shown interest. Mobasher (2007) express that drawback of this filtering technique is their tendency to over-specialize in suggesting the item to a user profile as user profiles are relayed on an attribute of the previous item opted by the user. Nevertheless, in the job domain, the job listed in the job board be available only for few days; due to the nature of the domain, the tendency to over-specialize in recommending the same item would not be any problem in the job domain recommender system. In domains like entertainment, user preference are tends to change depending on various factors, but In Job domain, the user tends to look for the job where he can use his previous skills. New recommendation of jobs can be made when there is a change in user preference, i.e. if a user thinks to change his/her job domain by updating his new skills and the job domain if he/she wishes. Another scenario of new recommendation is when new jobs are listed in the database; system would identify the properties of the job listed, such as

job domain and skills required for the job and matches with the users with a high similarity score.

Rule-based Filtering (RBF): These filtering techniques depend upon decision rules such as an automatic or manual decision rule that are manipulated to obtain a recommendation for the user profile. Currently, the E-commerce industry uses a rule-based filtering technique to recommend an item based on the demographic region of a user, purchase history, and other attributes that can be used to profile an user. A drawback in rule-based filtering is user feeds the information to the system. These inputs are utilized as a description of a user profile or can be considered as a preference of a user, defined by the user. Thus the data acquired is prone to bias. With the age of the user's profile, recommendation tends to hit the saturation and become staticMobasher (2007).

Hybrid filtering (HF): As the title describe, its incorporation of multiple techniques to improve the performance of recommendation. The previously discussed recommendation technique has its weakness and strengths. In order to get a better recommendation and overcome the challenges posed by earlier techniques, this technique is sought after. All of the learning/model-based techniques suffer from cold-start in one or other form. It is a problem related to handling a new user or new item. These and other shortcomings of the CF,CBF, and RBF could be resolved by using hybrid filtering techniques Burke (2007); Jain and Kakkar (2019); Dhameliya and Desai (2019).

The surveys conducted by Burke (2002) and Dhameliya and Desai (2019) have identified different types of hybrid filtering techniques that could be used by integrating CF, CBF, and RBF.

1. Weighted: The similarity score obtained from different recommendation components are coupled numerically to get one better recommendation.

2. Mixed: Recommendations obtained from different recommending techniques are put together and presented as one recommendation.

3. Switching: choosing one among the recommendation components based on the scenar- ios where it suits best.

4. Feature Combination: Attributes derived from diverse knowledge origins are fused and supplied to a recommendation algorithm.

5. Feature Augmentation: One recommendation technique is used to compute a set of attributes of user or item, which is then part of the input to the next recommenda- tion technique. Two or more recommendation techniques are serialised to get on recommendation.

6. Cascade: Recommending systems are given strict priority, with the lower priority ones breaking ties in the scoring of the higher ones. Here one Recsys technique refines recommendation of another.

There had been attempts to develop a recommendation system by several researchers. One such implementation was done by Rafter *et al.* (2000). They had devised a hybrid Recsys CASPER for Job finding search engine. They had implemented an automated collaborative filtering module and personalized case retrieval module in their job recommendation system. ACF module utilized user behavior information such as read time and activity on the page during his time on the system to profile the user. Similarity measure such as the Jaccard index and other clustering algorithms was used for similar grouping user against target user. Their other module PCR finds the similarity between the user's query and jobs in the system. The module computes similarity with a target user's query and jobs from the job case base using different similarity measures. This system has faced sparsity and scalability problems.

2.1 Natural language processing

These are the times that can be considered as an era of data. Every keystroke hit on twitter, online news, or in a research paper is recorded somewhere on the internet. All these generated data are available for the analysis through many means. In this abundance of data, Text data holds the majority of the share. Most of these text data are in an unstructured form. To put the abundance of text data into a perspective, a trillion-plus query per year is being handled by Google, and Whatsapp handles 30+ billion messages per day. That beingsaid, how do we extract information from the unstructured text data or how can we make machine understand what the text is about? To answer all the questions, Text analysis is a most sought after technique to extract useful information from the text data. Text analysis can be performed by utilizing techniques such as Natural language processing. Natural language processing is a process of information retrieval from unstructured data. It refers tothe utilization of computers to process natural language(Brants, 2003). The advancement in the personal assistant, text summarizing, and methods to caption a subject is due to the successful research in the field of NLP. Search engines like google and other industry leaders utilize NLP to its full extent. The gap between industry and academia in the field of NLP isvery minimal as there is an advancement in the NLP; the business has tried implementing and has brought closer to everyone's life.

In Recsys for the hiring domain, the data we handle here is nothing other than text data. A user profile describes the details about user experience and skills he/she familiar with. On the other hand, the job listed has information as job title, skills required to fulfill the role. All these information is filled with text data. In this scenario, we utilize the Natural Language Processing to measure the similarity between Jobs by checking the similarity between the job title and job description of the listed job. Determining the text-similarity is an essential task in several industrial application such as query search, text summarizing and video tagging(Kenter and De Rijke, 2015). In earlier studies, researchers have used

different approaches to identify similarity between the text by using edit distance algorithm which is discussed by Mihalcea *et al.* (2006), lexical overlapping technique (Jijkoun *et al.*, 2005) as this might work in most cases but can't rely on these technique because of its frail nature(Kenter and De Rijke, 2015). In such cases, we rely on technique called word embedding. This is huge development in the field of distributional semantics. As this requires only a large amount of unlabelled word data. These words are represented in semantic space as a vector. That is, words that are semantically similar will stay close in the semantic space.In order to retrieve terms that based on the similarity between two terms, we can utilize most well know method called word2vec a vector space model then we can use cosine similarity to measure the similarity between them (Shrestha, 2011; Barrón-Cedeno *et al.*, 2009). This model can also be used to determine similarity between the sentences(Barzilay and Elhadad, 2003). It's a group related model which is used to produce word embedding and these are set of language modelling and feature learning techniques of NLP where words are mapped to real values in the vector. Typically word2vec takes large set of words which is called corpus as a input and produces vector space with dimensions being in hundreds(Mikolov *et al.*, 2013). Once vector space model is generated we can use similarity measuring method to determine the distance or how similar is the word with which we are comparing. To find similarity in vector space we can use similarity measures like Cosine similarity and Jaccard similarity.

2.1.1 **Jaccard Coefficient**: Jaccard Coefficient is a method to compare elements of two sets to identify which elements are shared between two sets and which are distinct. It's similarity measure for two sets of data with result ranging from 0% to 100%. Two sets can be said similar, when result is close to 100% . Formula for Jaccard Index is as shown below(Sternitzke and Bergmann, 2009),

$$Jaccard\ Index = \frac{Number\ of\ Elements\ common\ in\ two\ sets}{Number\ of\ Elements\ in\ two\ sets} \qquad (2.1)$$

The above formula can be put into notation as below,

$$J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|} \tag{2.2}$$

2.1.1 Cosine similarity: Cosine similarity is also a measure to find similarity between two sets of non zero vector. It is a weighted vector space model utilized in the process of information retrieval. The similarity is measured by using euclidean cosine rule,i.e., by taking inner product space of two nonzero vector that measures the cosine of the angle between the two vectors. If the angle between two vectors is 0 deg , then the cosine of 0 is 1; Meaning that the two non zero vectors are similar to each other.In order to weight the words we have used the well-known word2vec vector space model(Rong, 2014; Herremans and Chuan, 2017).

$$Cosine\ Similarity(A,B) = cos(\theta) = \frac{\sum_{i=1}^{n} A_i * B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} * \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{2.3}$$

We can also compute cosine distance by using below equation,

$$Cosine\ Distance = 1 - (Cosine\ Similarity) \tag{2.}$$

2.2 Inferences

Based on all the research methodologies and techniques reviewed in this chapter, the CF technique cannot be considered as it does not satisfy the aims of the research. As the datasetof the user does not hold the information of rating against a particular job, we will not be ableto create a rating matrix that requires for CF technique. Instead, I have chosen to implement content-based filtering. I used multiple attributes in the user data to create a user profile and recommend the job to those profiles which have a high similarity score received from cosine similarity. Also, i have given higher weights to job skills when compared to the job domainof the user while computing similarity scores between user profile and job.