جـــامـعـــة الـــشــرق الأوســط
MIDDLE EAST UNIVERSITY

Latent Semantic Analysis (LSA) and ontological concepts to Support e-recruitment

التحليل الدلالي الكامن ومفاهيم الدلائل لدعم التوظيف الالكتروني

By

Khaleel Quftan Al-lasassmeh

Supervisor

Dr. Ahmad K. A. Kayed

Submitted in Partial Fulfillment of the Requirements for the Master Degree in Computer Science

Department of Computer Science

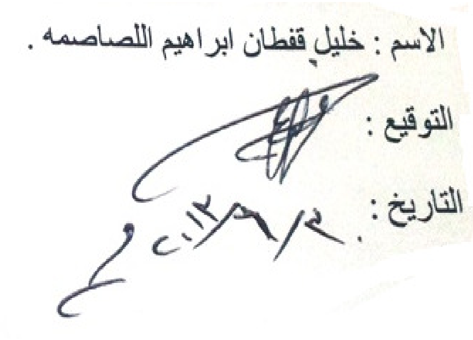
Faculty of Information Technology

Middle East University

June, 2013

تفويض

أنا **خليل قفطان ابراهيم اللصاصمه** أفوض جامعة الشرق الاوسط بتزويد نسخ رسالتي المعنونة بـ " التحليل الدلالي الكامن ومفاهيم الدلائل لدعم التوظيف الإلكتروني " للمكتبات الجامعية أو المؤسسات أو الهيئات أو الاشخاص المعنيين بالأبحاث والدراسات العلمية عند طلبها.



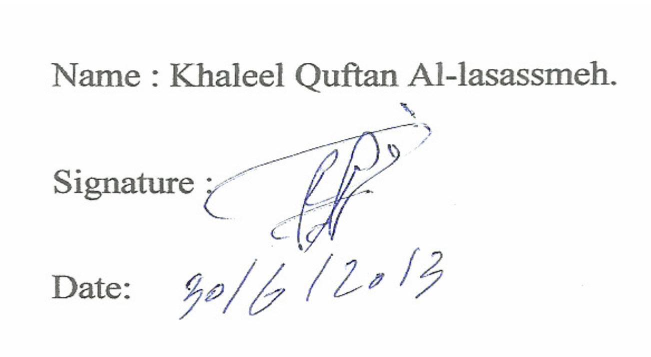
الاسم : خليل قفطان ابراهيم اللصاصمه .

التوقيع :

التاريخ : ٢/٨/٢٠١٢م

Authorization Form

I, The Undersigned (Khaleel Al- Lasassmeh), authorize the Middle East University for Graduate Studies to provide copies of my thesis to all and any university libraries and / or institutions or related parties interested in scientific researches upon their request .



Name : Khaleel Quftan Al-lasassmeh.

Signature :

Date: 30/6/2013

Discussion Committee Decision

This thesis has been discussed under the title " **Latent Semantic Analysis (LSA) and ontological concepts to Support e-recruitment** " this is to certify that the thesis entitled was successfully defended and approved: June 5 ,2013 .

Examination Committee Members **Signature**

1- **Dr. Ahmad K. A. Kayed** Supervisor ....

2- **Dr. Hiba Nassiredeen** Internal Member ....

3- **Dr. Jehad Al-Sadi** External Member..................

Acknowledgements

I would like to express my sincere thank to Dr. Ahmad Kayed for his continuous support, efforts, and dedication.

Dedication

I would like to express my thanks to my lovely parents who supported in my masters and in all academic stages of my life. They are the light in my path. To my brothers and family and patience for their love and support.

Abstract

Nowadays, supporting e-recruitment by new techniques and applications is one of the major research topics of applied semantic ontology. Most of the latest works, in this domain, has focused on analyzing the contents of CVs and matching them with job postings in order to find the most relevant CVs to a certain job posting. In this research, we present a new approach based on combining ontological concepts and latent semantic analysis together to match between CVs and job postings. This research investigates building a matrix in LSA that is based on using the ontological concepts and instances instead of words that are used in the traditional approaches. It also investigates enhancing the clustering using the concepts and instances. Our approach is better than other approaches in the sense that it takes into account the meanings and relationships among words, and also considers different words that have same meaning and words that are semantically equivalent. By implementing our methodology on a set of CVs and matching them with an actual job posting, and comparing our results with the results of other traditional approaches, our approach achieved 84% accuracy, which higher than all other approaches.

(الخلاصة)

في الوقت ألحاضر , دعم التوظيف الإلكتروني من خلال التقنيات والتطبيقات الجديدة هي واحدة من اهم الموضوعات البحثية الرئيسية لتطبيق الانطولوجيا الدلالي . معظم الاعمال في الوقت الحالي ركزت في هذا المجال على تحليل محتويات السير الذاتية ومطابقتها مع الاعلان الوظيفي , من أجل العثور على السير الذاتيه ذات اكبر صلة في العمل وفقا للإعلان الوظيفي . في هذا البحث نقدم نهج جديد يقوم على الجمع بين دلائل المفاهيم والتحليل الدلالي الكامن معاً للمطابقة بين السير الذاتية وإعلانات الوظائف. يتحرى هذا البحث بناء المصفوفة في تقنية التحليل الدلالي الكامن باستخدام دلائل المفاهيم وحالاتها بدلا من الكلمات التي تستخدم في النهج التقليدي . نهجنا حقق نتائج أفضل من النهوج الأخرى لأنه يأخذ بعين الاعتبار المعاني والعلاقات بين الكلمات ، و أيضا يشمل الكلمات المختلفة التي لها نفس المعنى والكلمات التي تعتبر متكافئة لغويا. من خلال تنفيذ نهجنا على مجموعة من السير الذاتية ومطابقتها مع الاعلانات الوظيفية ، ومقارنة نتائجنا مع نتائج النهوج التقليدية الأخرى، حقق نهجنا 84٪ نسبة دقه، وهذه أعلى نسبة على باقي الأساليب الأخرى.

TABLE OF CONTENTS

TITLE PAGE………………………………………………………………...……I

(تفويض)………………………………………………………………………...II

Authorization Form………………………………………………………..III

Discussion Committee Decision……………………………………………...IV

Acknowledgements V

Dedication VI

Abstract VII

(الخلاصة) VIII

TABLE OF CONTENTS VI

List of Tables...XI

List of Figures XII

Chapter One Introduction.. 1

1.1 Preface 1

1.2 Problem Definition.. 2

1.3 Contribution 3

1.4 Motivation 3

1.5 Objectives of the Thesis...3

1.6 Organization of the Thesis...4

Chapter Two ::Literature Survey 5

2.1 Theoretical Background 5

2.1.1 E-Recruitment..5

2.1.2 Ontology...6

2.1.3 Latent Semantic Analysis...7

2.1.4 Similarity Measure...10

2.1.5 Clustering...11

2.1.6 Tools used 14

Chapter Three ::The Proposed Model .... ….. ………20

3.1 CVMM Architecture. 20

3.2 Building the Ontology:...21

3.2.1 The Data Sets(sources of ontology)……………………………….. ...23

3.2.2 Extracting the concepts. ………………………………………….......23

3.2.3Pre-Processing the concepts……………….…………………….…..…24

3.2.4 The common concepts…………………………………………………25

3.2.5Instances of concepts:…………………………………………………….….26

3.3 Vector Generating (applying LSA) .... …27

3.3.1 Creating the Concepts -by-CVs Matrix: …………………....…………28

3.3.2 Apply a mathematical algorithms SVD ………………………………30

3.4 Finding Similarity To Match CVs And Job Posting 34

Chapter four :: Clustering of CVs 35

4.1 Clustering.. 35

4.2 Clustering of CVs and job posting : 37

4.3 Experiment result.. 38

4.3.1 Clustring based on concepts and Instances……………………….……39

4.3.2 Clustring by Tradational Approach. ……………..……………………..…40

Chapter five Experimental Result 42

5.1.Expert Judgment (the ground truth).. 43

5.2 - The similarity (keyword) :.. 46

5.3 COS- Word (traditional LSA) : 47

5.4- COS-(concepts only) : 48

5.5 CVMM (the proposed approach) :.. 49

5.6 Analysis of experiment result. 50

Chapter Six ::Conclusion and Future Work 52

6.1 Conclusion... 52

6.2 Recommendations and Future work………………………………….……..53

References……………………………………………………………………...….54

WEBSITES REFERENCES…………………………………………………...….58

APPENDICES……………………………………………………………………..59

List of Tables

Number	Title	Page
3-1	Sample of concepts frequency	26
3-2	Sample of data from matrix (concepts-by-CVs)	29
3-3	Unitary matrices U	31
3-4	Diagonal matrix	31
3-5	V^T which is conjugate transpose of V	32
3-6	The original matrix M'	33
4-1	The vectors of CVs	38
4-2	Experts clustering of CVs and job posting	39
4-3	Clustering based on concepts and instances	39
4-4	Result of CVS by traditional approach	41
5-1	Glossary of the approaches	42
5-2	Experts CVs and job posting matching	43
5-3	Result of matching (Keyword)	46
5-4	Result of matching Cos-word(Traditional LSA)	47
5-5	Result of matching Cos(concepts)	48
5-6	Result of matching by CVMM	49
5-7	Success rate for all approaches	50

List of Figures

Number	Title	Page
2-1	Factorization form for matrix M	9
2-2	Unitary matrix U	9
2-3	Cosine similarity equation	10
2-4	Euclidean similarity equation	11
2-5	Equation describe K- Means	13
3-1	Graphical description of CVMM	20
3-2	Building the ontology	22
3-3	OntoGen 2.0 interface	24
3-4	WordNet interface	27
3-5	Vector generation process	28
3-6	Equation to calculate M'	30
3-7	Cosine similarity equation	34
4-1	Data clustering	35
4-2	Data clustering process	37
4-3	Text clustering tool	40
5-1	CV12	44
5-2	CV7	45
5-3	Chart of result for all approaches	51

Chapter One
Introduction

1.1 Preface

Nowadays, many governmental services are being carried out via the Internet which makes them easily accessible and widespread; 90% of human resource managers in many countries consider the Internet an effective communication medium and rate the Internet as an important recruitment channel (Keim T., 2005). Examples of Internet services provided include: vacancy announcements, job postings, interviewing, filling online application forms, communicating with recruitment agencies and consultants, organizing appointments, conducting online tests, sending email notifications to applicants ... etc.

The recruitment process is a clear example of how automation and the utilization of a widespread medium such as the Internet enormously save cost and time. Automation also aids in choosing the appropriate applicant for the job by filtering a large number of CVs and identifying the most suitable applicant for the position. This thesis contributes to the automation of finding the appropriate applicant by developing a new approach based on Ontology and Latent Semantic Analysis (LSA). Ontology provides a set of concepts and their interrelationships in a specific domain that facilitates understanding and automatic processing of text (Zhang J. et al. 2012). According to Gruber (1993) "Ontology is a formal, explicit specification of a shared conceptualization"). LSA is an approach that uses a method called Singular Value Decomposition (SVD) to analyze and find the statistical relationships among words in large corpora of text. Our methodology uses semantic concepts and instances in LSA to create a matrix and use this matrix to compute SVD that is used later in the selection process.

This research also contributes to the automation by enhancing the clustering process by using concepts and instances instead of keywords. Clustering is a statistical technique used to organize collection of files or texts into meaningful clusters based on their similarity (Konda.S 2007).

1.2 Problem Definition

Evaluating CVs and determining the most suitable CV for a certain job is a tedious and time consuming task. Many methodologies have been devised to aid in the selecting and clustering of CVs that are best suited for a certain job posting. These methodologies and frameworks vary in their matching accuracy. Accuracy in such applications is always the subject of constant research with the aim of achieving better results and making more accurate judgment. Our research attempts to combine various techniques such as ontology and LSA in order to achieve more accurate results.

As far as we know there are no one have used ontology to build LSA, while there are many researches have used LSA to build ontology.

A challenge for this research is finding the most relevant concepts in the domain of CVs and job postings. This research addresses the following issues:

1- How to use ontology in building the LSA to enhance the matching process.

2- How to use the clustering technique that is dependent on ontology and LSA to enhance the CVs matching process.

3- The utilization of ontology concepts and LSA to support a more accurate decision making process.

1.3 Contribution

This thesis contributes in the following findings :

1- Using the ontology with LSA in the clustering process in the domain of CVs enhances the matching process.

2- Using instances along with ontological concepts enhances the matching process and produces more accurate results.

3- Explains how to use ontology to enhance the LSA.

1.4 Motivation

Enhancing the matching between CVs and job postings saves more time, cost and effort in governmental organizations. This saved time maybe allocated for other important and critical tasks. Also, improving the accuracy of choosing the suitable person for the job, also contributes to the overall performance of the organization, and also saves costs and money on the long run.

1.5 Objectives of the Thesis

The following are the objectives of this thesis :

1. To enhance the CVs matching process and to produce more accurate results.
2. To enhance the LSA by using ontological concepts and instances.
3. To enhance the clustering process by using ontological concepts.

1.6 Organization of the Thesis

Chapter 2 presents a theoretical background about the ontology, how build the Ontological concepts, latent semantic analysis, singular value decomposition, clustering, how the cosine similarity is measured, and the matching process between CVs and job posting.

Chapter 3 presents the proposed model and the process of applying the ontological concepts and LSA to filter the CVs according to the job posting. It also describes the details of our approach to compute the semantic similarity depending on the combination between the ontology and latent semantic analysis, and also presents the experimental results.

Chapter 4 presents the enhanced clustering process of the CVs and the result of enhanced clustering process depending on the ontology and LSA.

Finally Chapter 5 presents a discussion of the thesis, conclusion and future work .

Chapter Two:
Literature Survey

This chapter presents theoretical background about the ontology and the latent semantic analysis (LSA) technique. There are two main processes that are directly related to our work; the first is to build the ontology for a particular domain and the second is using the LSA technique to represent the data.

2.1 Theoretical Background

This section presents the necessary theoretical background for understanding the topics related to the thesis.

2.1.1 E-Recruitment

According to Mochol, M.,et al (2004) e-recruitment services are part of e-government services that uses the technology for better performance and efficiency. Each domain of work has its own different job posting specifications that depends on the market and the educational needs for the job. An organization needs to provide the full details in the job posting to describe the job.

The e-recruitment process in an organization have four main phases:

1. Describing the requirements of the job position.

2 .Publishing the job posting.

3. Receiving of applications.

4. Decision making.

Nowadays, applicants can look for a job in any web site that provide e-recruitment services. Also, they can use the search engine to find a job. Applicants must submit a CV (Curriculum Vitae) to apply for a certain job posting. Additional information in the applicant CV gives more opportunities for applicants to provide more information about their qualification. The advantage of CVs over filling an application form is that applicants have a better opportunity to express themselves in their own way due to the absence of the limitations that are enforced by the application forms. In the application forms approach, candidates fill their information in boxes and predefined templates. However, absence of the limitations in the CVs make CVs harder to assess consistently. The evaluation of all received CVs in e-recruitment becomes a long and tedious process and takes time, effort and money especially when the number of applicants is very high.

2.1.2 Ontology

The term 'ontology' is derived from the Greek words 'onto' which means being, and 'logia' which means written or spoken discourse. According to Gruber (1993), ontology is an explicit specification of a conceptualization. Gruber (2007) defined it as a set of representational primitives that include classes, attributes and relationships that are used to model a domain of knowledge or discourse. Examples of classes include; sets, collections, or types of objects(person, animal, food, table, etc), examples of attributes include; properties, features, characteristics, examples of parameters that objects can have and share include; a Person class has the properties of gender, height, weight, hair color , mobile no ... etc, examples of relations include; Khaleel lives in karak and Karak is located in Amman.

Deploying such ontologies in the domain of matching between CVs and job posting will enhance the selection in e-recruitment. Ontology provides a good method to understand a domain of interest to support communication between humans and computer (Maedche and Staab, 2001).

The first step in building a new ontology is identifying the domain of the ontology and why the ontology is being built, the intended users. The second step is determining and specifying the sources which might be; documents, experts and existing ontologies. The third step is the actual building of the ontology using a suitable ontology building tool such as KAON, Protégé, etc. (Fernández-López M. 1999)(Bermejo J. 2007).

In the ontology building process, according to Kayed(2010), extracting ontology concepts consists of concepts that are not only the most frequent terms, but also those that have high ontological relevance keywords. The ontology building process, according to Kayed(2010), is extracting the ontological concepts that consist of concepts that are not only the most frequent terms, but also those that have high ontological relevance keywords.

2.1.3 Latent Semantic Analysis

LSA is an intelligent information retrieval technique that analyses collection of text to find the semantic meaning among the documents using mathematical algorithms. LSA is a highly parameterized statistical method, and its effectiveness is driven by the setting of its parameters which are set differently based on the task (Cosma G. and Joy S, 2012).

LSA is also a statistical approach that analyzes the statistical relationships among words in a large accumulated text by using a method called (SVD) to find the global knowledge indirectly from local co-occurrence data in a large body of representative text. LSA finds a projection matrix that converts the high dimensional vector space representations of documents to a lower dimensional space built with latent factors (Deerwester S. et al, 1990).

The application of LSA for information retrieval in literature dates back to 1988. LSA is also known as Latent Semantic Indexing (LSI), and the term LSI is used for tasks concerning the indexing or retrieval of information, whereas the term LSA is used for tasks concerned with analyzing texts such as automatic essay grading and text summarization. (Cosma, G., & Joy, M. (2012).

We used the LSA as a potential technique to support the e-recruitment. because of the variety in the words used to describe the CVs and job postings (Variety in the words that people use to describe the same thing (synonyms)), CVs and job posting matching methods often fail to retrieve the information that is used to measure the real similarity. Empirical evidence suggests that the likelihood of two people choosing the same keyword to describe a familiar object or concept is between 10% and 15% (Furnas, G. W., eta (1984). Furthermore, each word may be have more than one meaning (polysemy), which leads to irrelevant information being retrieved. From this perspective, exact lexical-matching methods are deficient for information retrieval(Dumais, S. T., eta(1988))

The LSA technique is comprised of mathematical algorithms that are applied to text collections. Initially a text collection is pre-processed and represented as a term-by-documents matrix containing terms and their frequency counts in files.

SVD decomposes this term-by-document matrix into separate matrices that capture the similarity between terms and between documents across various dimensions in space. These matrices called U, **Σ and** V*.

The SVD of any matrix M is a factorization of the form:

$$M = U\Sigma V^*,$$

Figure 2-1: Factorization form for matrix M**.**

where U is a m×m real unitary matrix where a matrix U is unitary if:

$$U^*U = UU^* = I$$

Figure 2-2: Unitary matrix U.

where **I** is the identity matrix and **U** * is the conjugate transpose of **U**. And **Σ** is an m×n rectangular diagonal matrix (matrix in which the entries outside the main diagonal) (are all zero) and V* (the conjugate transpose of V) is an n×n real or complex unitary matrix. (Golub and Van Loan 1996).

The aim is to represent the relationships between terms in a reduced dimensional space such that noise (i.e. variability in word usage) is removed from the data and therefore uncovering the important relations between terms and documents obscured by noise [Berry, M. W.,eta. (1995)). LSA aims to find the underlying (latent) relationships between different terms that have the same meaning but never occur in the same document.

2.1.4 Similarity Measure

The similarity measure reflects the degree of closeness or separation between objects. Choosing of similarity measure is crucial to find similarity, especially for a particular type of documents matching and clustering algorithms. Distance measures represents the distance or similarity between the two objects as a single numeric value. There are several similarity measures to calculate distance or similarity of a text document. Huang, A. (2008).

Cosine Similarity

Cosine similarity is one of the well known similarity measures applied to text documents, such as in numerous information. When two documents are represented as term vectors, the similarity of two these documents corresponds to the correlation between the vectors. This is equal the cosine of the angle between vectors.

Given two document t_a and t_b their cosine similarity is

$$\boxed{SIM_C(\vec{t_a}, \vec{t_b}) = \frac{\vec{t_a} \cdot \vec{t_b}}{|\vec{t_a}| \times |\vec{t_b}|},}$$

Figure 2-3: Cosine similarity equation .

Where t_a and t_b : m - dimensional vectors over the term set T = {t1, . . . , tm}. Each dimension represents a term with its frequency in the document, which is non-negative. As a result, the cosine similarity is non-negative and bounded between [0,1].

Euclidean Distance

Euclidean similarity is a standard metric for geometrical problems. It is the ordinary distance between two points and can be easily measured with a ruler in two or three dimensional space. It is widely used in clustering problems, including text clustering. It is also the default similarity measure used with the K-means algorithm.

Measuring similarity between text documents, given two documents, the Euclidean distance of the two documents is defined as:

$$D_E(\overrightarrow{t_a}, \overrightarrow{t_b}) = (\sum_{t=1}^{m} |w_{t,a} - w_{t,b}|^2)^{1/2}$$

Figure 2-4: Euclidean similarity equation .

where the term set is T = {t1, . . . , tm}.

2.1.5 Clustering

Clustering is a statistical technique used to build many classes having similar properties, with each class containing a set of details that have a high degree of association among other members of the same class (Anderberg, M. R. (1973). Clustering algorithms can be classified according to:

- The type of data that is passed to the clustering algorithm.
- The clustering criterion defining the similarity between data points.
- The theory and fundamental concepts on which clustering analysis techniques are based.

According to the method and technique used to define clusters, the algorithms can be broadly classified into the following types (Jain et al., 1999):

- Partitional clustering attempts to directly decompose the collection of text with N objects into M clusters such that no overlap is allowed, and each item is contained in the most similar cluster to it . More specifically, they attempt to determine an integer number of partitions that optimize managing of data.
- Hierarchical clustering produces a nested data set like a tree, where each pairs of documents or clusters are successively linked together until every cluster in the data set is connected. This is performed by merging smaller clusters into larger ones, or by splitting larger clusters into smaller ones. The result of the algorithm is a tree of clusters.
- The key idea of Density-based clustering is to group neighboring objects of a data set into clusters based on density conditions.
- Grid-based clustering. This type of algorithms is mainly proposed for spatial data mining. Their main characteristic is that they quantize the space into a finite number of cells and then they do all operations on the quantized space.

According to Mount, D. (2005) in Partitional clustering category, K-Means is a commonly used algorithm.

The aim of K-Means clustering is the optimization of an objective function that is described by the equation:

$$E = \sum_{i=1}^{c} \sum_{x \in C_i} d(x, m_i)$$

Figure 2-5: Equation describe K-Means clustering.

Where *mi* is the center of cluster *Ci*, while d(*x*,*mi*) is the Euclidean distance between a point *x* and *mi*. Thus, the criterion function attempts to find the distance between each point from to the center of the cluster to which the point belongs.

K-Means algorithm is composed of the following step:

1 Initializing a set of c cluster centers.

2 Assigns each object of the data set to the cluster whose center is the nearest, and recomputes the centers.

3 The process continues until the centers of the clusters stop changing.

The main object of clustering is to separate a collection of unlabeled data set into a groups , rather than provide an accurate characterization of unobserved samples Baraldi (2002).

2.1.6 Tools used .

MATLAB is a high-level language and has an interactive environment tools for numerical computation , visualization ,and programming. MatLab can be used to analyze data, develop algorithms, create models and applications, and build tools. It has built-in math functions that enables you to explore multiple approaches and reach a solution faster than with spreadsheets of traditional programming languages, such as C/C++ or Java.[1]

OntoGen is an ontology editor focusing on editing of topics that are connected with each other in different types of relations . It has two components which are, the semi-automatic component and the data-driven component. The semi-automatic component is an interactive tool that aids the user during the ontology construction process. It suggests: concepts, relations between the concepts, names for the concepts, provides a good overview of the ontology to the user through concept browsing and various kinds of visualizations. The data-driven component is based on the underlying data provided by the user typically at the beginning of the ontology construction. The data reflects the structure of the domain for which the user is building ontology. OntoGen system combines text-mining techniques with an efficient user interface to reduce both: the time spent and complexity for the user. It is chosen because it provides better flexibility for meta-modeling, enables the construction of domain ontologies; customize data entry forms to enter data.

[1] .(http://www.mathworks.com/products/matlab/)

WordNet is a machine readable dictionary. It is a project created at the Cognitive Science Laboratory at Princeton University as a measuring of Semantic Relatedness. This project contains only open-class words (nouns, verbs, adjectives, and adverbs).It does not contain closed-class words such as (pronouns, conjunctions, and prepositions). WordNet groups sets of synonymous word instances into synonym sets.(Fellbaum, C. (2010)).

Related Work

In 1996, two longtime friends and former bankers, Robert Ruff and Jeffrey Smith, started Sovren Group which began as a staffing business with a focus on the financial and accounting markets. The company's sole business is providing the best parsing and matching components for recruitment intelligence and providing many services such as Resume/CV Parsing and Semantic Matching solutions. Sovren offers two product lines: The Sovren Resume/CV Parser (Quickly extract a wealth of information from resumes and CVs in any format), and The Semantic Matching Engine (SME)(find the best matching candidates and jobs with this sophisticated profile matching engine)[2] (Robert H.et al).

(Landauer, K. et al ,1997), established novel theory of acquired similarity and knowledge representation . By using global knowledge indirectly from local co-occurrence data in a large body of representative text , the approach is based solely on a general mathematical learning method that achieves powerful inductive effects by extracting the right number of dimensions to represent objects and contexts. Relations to other theories, phenomena, and problems are also sketched.

[2] http://www.sovren.com

(Mochol, M. et al 2004) proposed the application of semantic Web technologies in the recruitment domain. It used existing classifications and standards for ontology development in human resources that gived semantic annotation of job postings and job applications using controlled vocabularies, in contrast to the free-text descriptions, to give better results and improve query results to deliver an ordered list of the best candidates for certain job matching.

(Bizer, C. et al ,2005), described how online recruitment processes can be streamlined using semantic web technologies. It also described prototypical implementation of the required technological infrastructure by focusing on the ontologies, the data integration infrastructure, and on semantic matching services. They also analyzed the potential economic impacts of the realization of the scenario from the perspectives of the different participants.

(Lv, H. et al 2006) proposed an intelligent recruitment platform with a skill ontology-based semantic model and its matching algorithm, by which recruitment websites can automatically find the most appropriate job-seekers for a certain job post and positions for a job-seeker. The main idea of the research is converting the skill matching process into seeking the shortest path between two concepts vertexes in certain sub-ontology, and the total of edges' weights along the shortest path measures the similarity between them. The smaller the weights total are, the more similar the skills are semantically. The research supposed that its framework of semantic matching can be applied to some other fields such as computer aided industry design and manufacturing information engineering.

(Yahiaoui, L. et al ,2006), proposed a scenario for automatic e-recruitment through the web in which the applications and job descriptions are matched through semantic annotation and indexing based on competency profiles. It also used the semantic matching computation coefficient as a matching algorithm.

(Mochol, M. et al ,2007), proposed an e-recruitment methodology based on semantic web. They adopted the ontology from knowledge nets. They proposed an approach of query approximation and showed how it is applied to the semantic job portal, which improved the results of job search.

(East W., 2008) provided a justification for the use of e-Government applications for public facility delivery projects. The costs and benefits from using such a system in public practice in the US described. The emerging impacts of e-Government systems on the facility acquisition community are also identified.

(Bettahar F., 2009), proposed a project that focused on the semantic requirements of governments at local, intermediate and regional levels, that are needed to build flexible and interoperable tools to support the change towards e-Government services. They proposed an ontology to represent knowledge and to achieve the required level of semantic interoperability. The key feature of the system is a unique and multimodal ontology that are used simultaneously. This key features enables describing domain knowledge, adding semantics to agency services, indexing various documents in knowledge bases used by civil servants, and supporting the interaction between the users and the system. The research also presented the challenges of using ontology in e-government environments, such as the lack of expressivity of the formalism chosen for interoperability in the project and the risk of inconsistency when the ontology changes.

(Fazel-Zarandi, M. et al 2009), proposed an ontology-based hybrid approach to efficiently match job seekers and job descriptions. Their research used node-based semantic similarity measure which is a terminological matching technique.

(Lv, H., et al ,2010) proposed a new semantic similarity methodology that is based on the distance of a concept tree structure which is both terminological and structural. This approach focused on electronic job recruitment process by analyzing the weight setting methods of a demand index by AHP. This creates a weighted sub-ontology draw, which represents all special field skill concepts and their relationship.

(Amdouni, S.et al ,2010) proposed a framework that represents an automatic tool of CV analysis for purpose of normalizing the CV content according to the structure adopted by Europass CV. The CVs of this research was in French language.

(Kayed a.et al. 2010) developed a new ontology in the domain of requirements engineering process for E-gov applications. The research provided common concepts and understanding of the requirements for many E-gov applications. This new ontology and concepts aids and enables software engineers to find out common concepts to describe requirements for different domain models used in developing E-gov applications.

(Lintean, M., et al. ,2010) investigated the impact of weighting schemes on LSA's ability to capture semantic similarity between two texts. They worked with texts varying in size from sentences to paragraphs. The conducted experiments revealed that for sentence-level texts, a combination of type frequency local weighting in combination with either IDF or binary global weighting works best. For paragraph-level texts, a log-type local weighting in combination with binary global weighting works best. We also found that global weights have a greater impact for sententence -level similarity as the local weight is undermined by the small size of such texts.

(Vincent J. et al, 2011) proposed an approach named "Unscholed and King ontology" is applied to develop semantic ontology models in the government service domain. Firstly, the approach is applied to build a government domain ontology. Secondly, the domain ontology is evaluated for semantic consistency using its semi-formal representation in Description Logic. Thirdly, an alignment of the domain ontology with the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) upper level ontology is drawn to allow its wider visibility and to facilitate its integration with existing metadata standard. Finally, the domain ontology is formally written in Web Ontology Language (OWL) to enable its automatic processing by computers.

(Kayed, A. 2011), demonstrated several experiments to extract concepts to build ontologies that improved the description process for software components embedded in a web document. This paper built a new ontology (mainly concepts) for some software components that are then utilized to solve some semantic problems. The research collected many documents that describe components in .Net and Java from several and different resources. Concepts were extracted and used to decide which domain of any given description (semantic) is close or belong to.

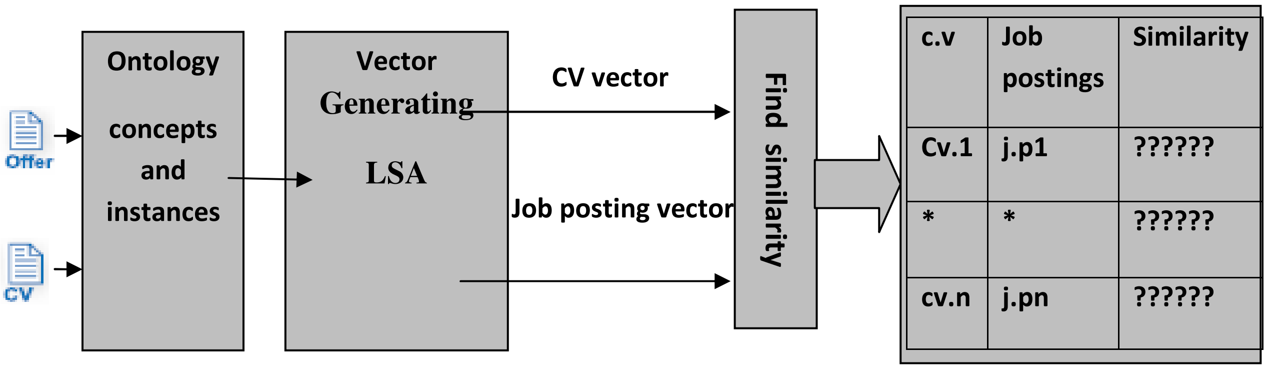
(Stamatios, A. et al, 2012) investigated and researched the possibilities applying Semantic Web and the technologies that support it. Their findings was that the semantic web is expected to provide effective solutions concerning a better exploitation of the information offered as well as producing and managing knowledge in the field of e-Government.

Chapter Three:
The Proposed Model

This chapter presents our model for matching CVs with job postings which is combined of two techniques; ontology and LSA. Our model will be referred to as CVMM: CV Matching Model. This chapter explains the architecture of our proposed CVMM model and presents the process of building the ontology in the domain of e-recruitment and also presents the process of applying the LSA technique and computing the semantic similarity between CVs and job postings using the cosine similarity. Building the ontology and applying the LSA are the two core concepts in our model.

3.1 CVMM Architecture.

CVMM includes three main phases which are building the ontology, the generation of vector(applying LSA), and finding the semantic similarity by applying the cosine similarity. This Architecture is illustrated in Figure 3-1.



c.v	Job postings	Similarity
Cv.1	j.p1	??????
*	*	??????
cv.n	j.pn	??????

Figure 3-1: Graphical description of CVMM.

3.2 Building the Ontology:

Building the ontology in CVMM focuses on concepts and instances to give a clear representation of all the words in the CV, without the need to use all the words. This allows an applicant to freely express information about himself and his qualifications starting from his education, experience, personal information and other interests. CVMM approach does not impose any restrictions or constraints on the format of the input CV or job posting that makes any CV in free text form acceptable.

In our work we used the teaching job postings in universities as the domain because it is widespread in e-recruitment. The following steps details our CVMM implementation:

1. Collecting CVs and job posting: we have used CVs for professors from the university of Jordan and job postings from multiple e-recruitment web sites (Bayt website and other similar sites)

2. Extracting the ontological concepts: Building of ontology requires the use of ontology building tools such as OntoGen and KAON. This thesis uses OntoGen to build the ontology.

3. Finding the common concepts : We extracted the most frequent concepts from the concepts in step (2). This has been performed by developing a small program in C++ to develop the count.

4. Pre-processing the common concepts: Spelling Correction, Separating concepts containing multiple words. This is necessary for the next step as the WORDNET accept only correct words and single word concepts.

5. Finding the instances of concepts: This thesis used the WORDNET too to extract the instance. The WORDNED has a database of concepts and their instances. This step of the research where we utilize the existing ontology.
6. Saving the processed concepts and their instances in a file to be used later.

These steps are illustrated in Figure 3- 2.

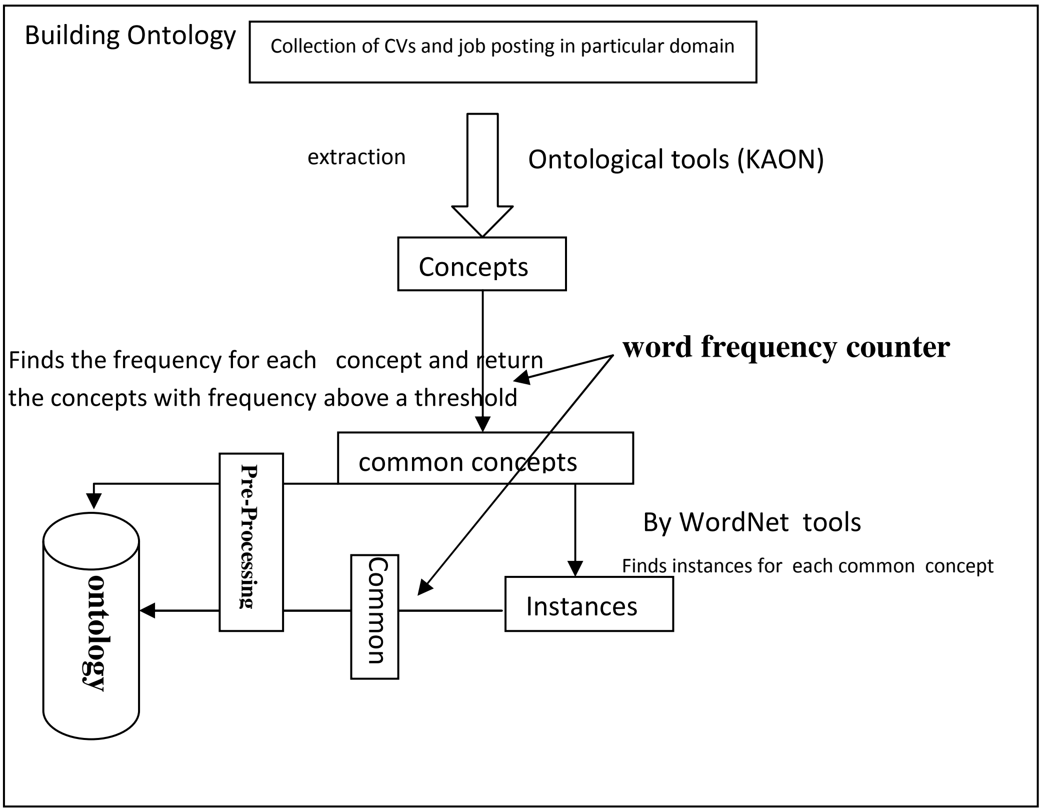

Building Ontology
Collection of CVs and job posting in particular domain
extraction
Ontological tools (KAON)
Concepts
Finds the frequency for each concept and return the concepts with frequency above a threshold
word frequency counter
common concepts
By WordNet tools
Finds instances for each common concept
Instances
Common
Pre-Processing
ontology

Figure 3- 2 : Building the ontology.

3.2.1 The Data Sets(sources of ontology)

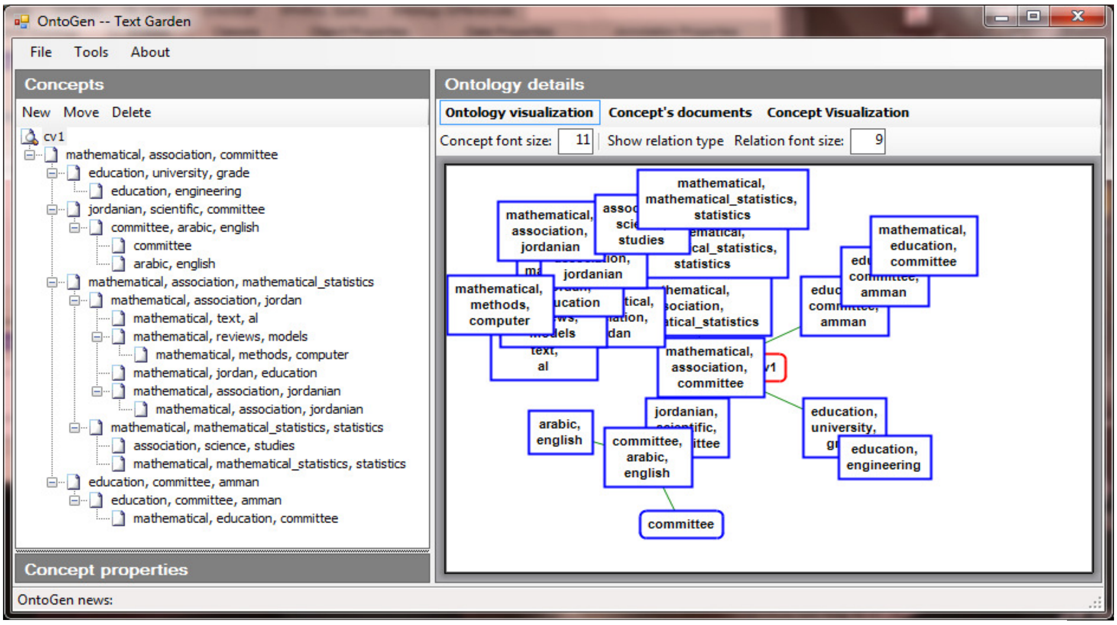
The sources or datasets needed for this approach are collections of CVs and job posting files. We chose 44 CVs and 5 job posting in the teaching staff domain. These C.Vs and job postings don't abide to fixed or predefined form and are widespread in e-recruitment. The CVs and job postings were taken from diverse fields such as, chemistry, mathematics, physics, engineering, animals production. These collected CVs are of real employees of the teaching staff in the university of Jordan. The job posting were taken from official employment sites such as (Bayt website).

3.2.2 Extracting the concepts.

After identifying the domain the sources of the ontology, the next step to extract the concepts. This is performed by using ontology building tools such as OntoGen. Ontological tools to extract a concepts from accumulated files. These concepts will be extracted from the datasets by ontology building tools such as OntoGen2.0. OntoGen is a semi-automatic and data-driven ontology editor focusing on editing of topic ontologies[3]. In OntoGen 2.0 tools, we can copy and extract the concepts and export them into a text file. Other ontology building tools do not offer the feature of exporting and copying the extracted concepts.

[3] (http://ontogen.ijs.si/)

In Figure 3-3, shows a screenshot of OntoGen 2.0 interface and it shows the extracted concepts for CV$_1$ in the left part of tool, and right part show the visualization of relations between the concepts.



```
cv1
mathematical, association, committee
    education, university, grade
        education, engineering
    jordanian, scientific, committee
        committee, arabic, english
            committee
            arabic, english
    mathematical, association, mathematical_statistics
        mathematical, association, jordan
            mathematical, text, al
            mathematical, reviews, models
                mathematical, methods, computer
            mathematical, jordan, education
            mathematical, association, jordanian
                mathematical, association, jordanian
        mathematical, mathematical_statistics, statistics
            association, science, studies
            mathematical, mathematical_statistics, statistics
    education, committee, amman
        education, committee, amman
            mathematical, education, committee
```

Concept font size: 11 Relation font size: 9

Figure 3-3 : OntoGen 2.0 interface .

The concept that extracted from the all CVs and job posting in accumulated data, will save in file named concepts to be used in the next step which is pre-processing the concepts to find a common concepts by calculating the frequency for each concept in all the file.

3.2.3 Pre-Processing the concepts

Concepts must be pre-Processed to be ready to be passed to the next steps; so that only concepts that have correct spelling will remain in the concept data file. Specialized programs might be used for this step, but in our case the data set was small enough that we could use MSWord to spell check the data.

The following is a list of pre-processing steps:

1. Spelling Correction.

2. Separating concepts containing multiple words .

3. Removing single letter words and stopping words such "at", "or", "is" ... etc.

The following example shows concepts after and before processing. The concept before processing is "{academic_as, assistant_professor ,committee_ph ,of, research_1, production_ effect}" and the concept after processing is "{academic , assistant , professor ,committee , research, production , ph, as, effect}".

3.2.4 The common concepts

The common concepts are the concepts that everyone agreed to use in writing

his CV. These concepts will have high frequency because they are repeated in more than one CV at least, and any concepts with high frequency are considered to be common concepts.

To find the frequency we used a custom word counter program written in C++. The threshold frequency for a common concept was set to "2" which means that each concept must appear in two CV files to be classified as a common concept. By applying word counter program on the file of concepts from the previous stage we got 90 concepts as common concepts and these common concepts.

Table 3-1 presents the most common concepts that appeared in the CVs and job postings. All common concepts was saved in file called common

concepts to be used in next step.

Table3-1 shows a sample of concepts with frequency.

Concept	Frequency	Concept	Frequency
research	**126**	**academic**	**12**
teaching	**44**	**Conference**	**48**
related	**41**	**committee**	**77**
scientific	**41**	**production**	**77**
training	**41**	**effects**	**71**
workshop	**39**	**education**	**23**
university	**34**	**meeting**	**23**

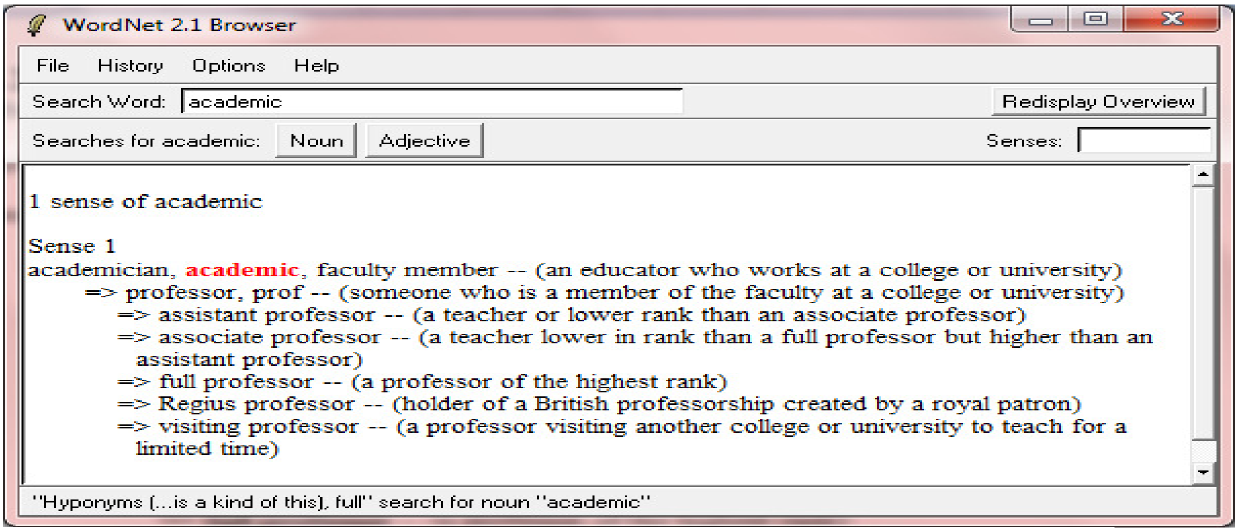
Table 3-1:sample of concepts with frequency.

3.2.5 Instances of concepts:

Instance is an individual ground-level object of a concept, and some concepts have instances that are used more frequently than other instances of the same concept. Also some instances are used more frequently than its concept. For example, the concept "academic" have many instances such as "assistant professor" and "associate professor".

To find the instances for common concepts we used a machine-readable dictionary, such as WordNet[3], which is a large lexical database of English nouns, verbs, adjectives and adverbs grouped into groups of cognitive synonyms. synonyms. For more details see http://wordnet.princeton.edu/.

Figure 3-4 Illustrates the instances of the concept "academic".



1 sense of academic

Sense 1
academician, academic, faculty member -- (an educator who works at a college or university)
=> professor, prof -- (someone who is a member of the faculty at a college or university)
=> assistant professor -- (a teacher or lower rank than an associate professor)
=> associate professor -- (a teacher lower in rank than a full professor but higher than an assistant professor)
=> full professor -- (a professor of the highest rank)
=> Regius professor -- (holder of a British professorship created by a royal patron)
=> visiting professor -- (a professor visiting another college or university to teach for a limited time)

"Hyponyms (...is a kind of this), full" search for noun "academic"

Figure 3-4 : WordNet interface.

Our data set produced more than 200 instances, and these instances were filtered were by the same threshold (minimum frequency of two instanced). The filtration produced 32 common instances.

3.2.6 Vector Generation (applying LSA):

In this step a vector is generated for each CV and for each job posting by applying the LSA. The generated vector represents the frequency of concepts and instances that are contained in the relevant CV or job posting.

The LSA Vector Generation process involves the following steps, and this process is illustrated in (3-5).

1- Creating a matrix of CVs and job postings by transforming the data in CVs and job postings into a m×n matrix such that m is the number of CVs and job postings, and n is the number of concepts and instances.

2- Applying the SVD (Singular Value Decomposition). The result of Appling the SVD to the matrix produced by step(1) is a new matrix that contains a vector representation for each CV and job posting.

This processes is illustrated in **Figure 3-5**

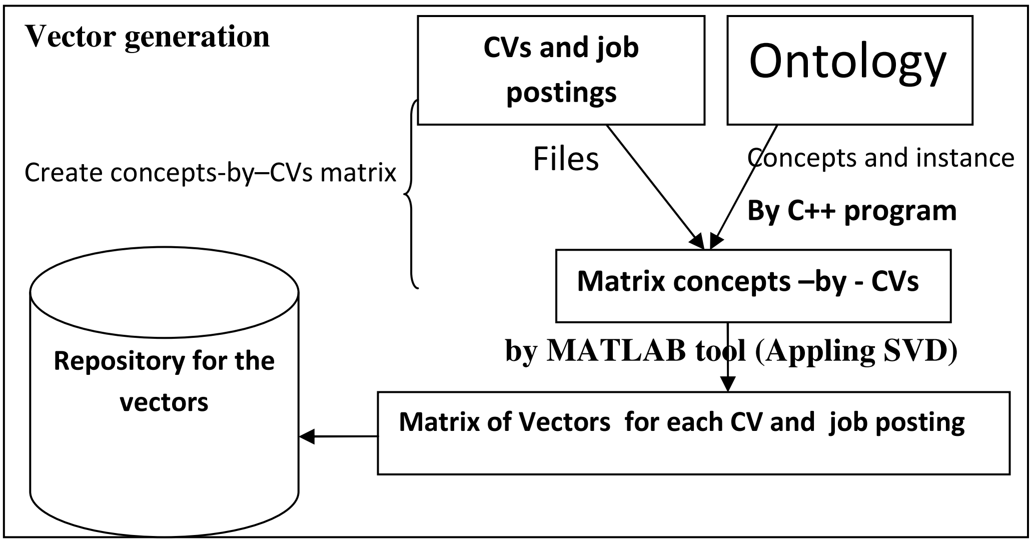

Vector generation
CVs and job postings
Ontology
Create concepts-by–CVs matrix
Files
Concepts and instance
By C++ program
Matrix concepts –by - CVs
by MATLAB tool (Appling SVD)
Repository for the vectors
Matrix of Vectors for each CV and job posting

Figure 3-5 : Vector generation process.

3.3.1 Creating the Concepts -by-CVs Matrix:

The first step in applying LSA is to transform all the data in job postings and CVs into a two dimensional matrix. The matrix has M rows and N columns, which each row Mi representing a CV or job posting vector, and each column Nj representing a concept or instance vector. A cell(i,j) contains the frequency of concept(j) in cv(i).

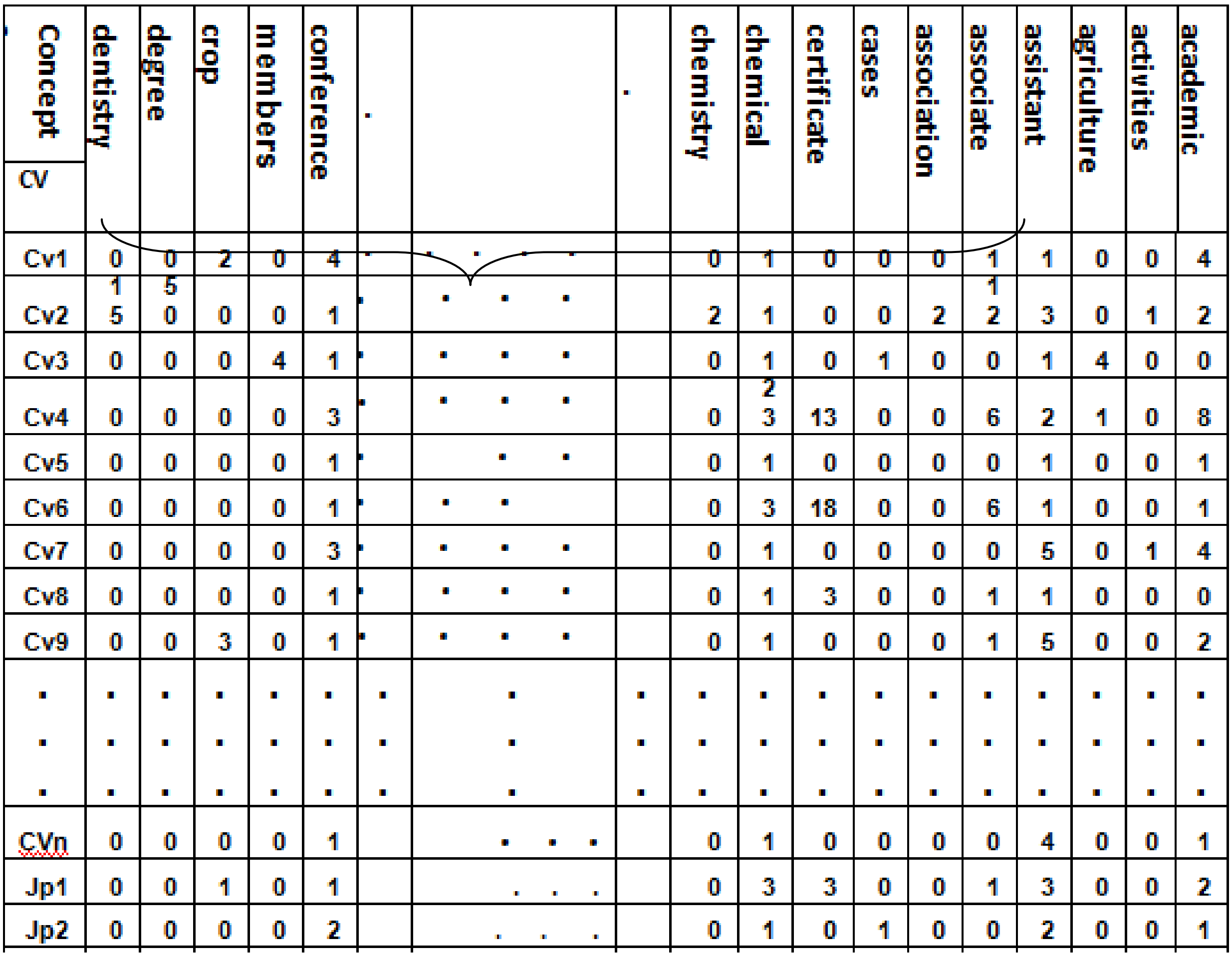
The matrix was generated with aid of a software application we developed and named it " word frequency counter ". This program accepts the concepts, instances, CV files, job posting files as inputs and produces the matrix as

output.

Table (3-2) shows an sample of the original matrix. The first row in the matrix has all the concepts and the instances from the ontology(dentistry, degree, crop, members). The first column contains the name of CVs and job postings. As a clarification, the element in cell A[1,2] has the values of 2, which is the frequency of concept (degree) in CV1 and cell A[3,5]= 1,which is the frequency of the concept "conferences" in CV3.

Also, each row M represents a CV as a vector and an example is the vector for CV1 ,which is the listing of counts of all frequencies in this CV.

The vector for CV1 is : "{0 0 2 0 4 6 0 .... 0 0 0 01 12 0........0 0 4 }".Alco, each column N represents a concepts vector. An example of a concepts vector is the vector of C1 which is "{0 15 0 0 0 0 0 0 0 0 0 0 0 0 ....0 0 0 0 }".



Concept CV	dentistry	degree	crop	members	conference	.		.	chemistry	chemical	certificate	cases	association	associate	assistant	agriculture	activities	academic
Cv1	0	0	2	0	4	.	. . . .		0	1	0	0	0	1	1	0	0	4
Cv2	15	50	0	0	1	.	. . .		2	1	0	0	2	12	3	0	1	2
Cv3	0	0	0	4	1	.	. . .		0	1	0	1	0	0	1	4	0	0
Cv4	0	0	0	0	3	.	. . .		0	23	13	0	0	6	2	1	0	8
Cv5	0	0	0	0	1	.	. .		0	1	0	0	0	0	1	0	0	1
Cv6	0	0	0	0	1	.	. .		0	3	18	0	0	6	1	0	0	1
Cv7	0	0	0	0	3	.	. . .		0	1	0	0	0	0	5	0	1	4
Cv8	0	0	0	0	1	.	. . .		0	1	3	0	0	1	1	0	0	0
Cv9	0	0	3	0	1	.	. . .		0	1	0	0	0	1	5	0	0	2
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
CVn	0	0	0	0	1		. . .		0	1	0	0	0	0	4	0	0	1
Jp1	0	0	1	0	1		. . .		0	3	3	0	0	1	3	0	0	2
Jp2	0	0	0	0	2		. . .		0	1	0	1	0	0	2	0	0	1

Table 3.2 :sample of data from matrix (concepts _by_CVs).

3.3.2 Apply a mathematical algorithms SVD.

The basic ideas of applying SVD is to take a high dimensional , highly variable set of data and reducing it to a lower dimensional space that exposes the substructure of the original data more clearly and orders it from most variation to the least. (Baker, K. (2005)).

The result of applying SVD will decompose the original Matrix into three other matrices (an illustrative fragment of them is shown in Tables 3.3, 3.4, and 3.5.):

$$U_{nxn} \;,\; S_{nxp} \;,\; V^{T}_{pxp}$$

where **U** and **V** are unitary matrices, V^T is the conjugate transpose of V, and S is a diagonal matrix. (see 2.3). Then, the matrix **M'** (the vectors-matrix) is calculated from:

$$\boxed{M' = U_{nxn} \text{ x } S_{nxp} \text{ x } V^{T}_{pxp}}$$

Figure 3-6: Equation to calculate M'

M' is the matrix that contains a vector for each CV and job posting.

All matrix operations are conveniently performed by MATLAB.

Table 3.3 : Unitary matrices U.

-0.071	0.837	-0.266	0.13	0.232	-0.198
-0.456	-0.295	0.216	0.613	0.378	-0.258
0.014	0.004	-0.025	0.001	-0.003	-0.061
0.065	0.048	0.23	-0.187	0.126	-0.234
0.015	0.011	0.014	-0.003	0.025	-0.033
0.069	0.066	0.195	-0.032	0.083	-0.213
-0.008	0.139	-0.051	0.02	0.095	-0.079
-0.071	-0.216	-0.319	-0.227	0.157	-0.065
-0.017	0.027	-0.036	0.008	0.072	-0.066
-0.454	0.107	0.393	-0.416	-0.055	-0.168
0.001	0.044	-0.016	0.004	0.068	-0.056
-0.041	-0.107	-0.255	-0.159	0.135	-0.07
-0.142	0.039	0.042	-0.081	0.01	-0.066
0.01	0.012	0.008	-0.006	0.034	-0.053
0.008	0.012	0.003	-0.013	0.036	-0.064
0.075	-0.017	0.038	-0.043	0.071	-0.094
0.253	-0.046	0.051	0	0.057	-0.199
-0.284	-0.041	-0.171	0.095	-0.438	-0.247
0.12	-0.05	-0.077	-0.007	-0.114	-0.137
-0.011	-0.006	-0.015	0.021	-0.014	-0.058
0.09	-0.019	-0.085	0.047	-0.176	-0.217
-0.049	0.012	-0.05	-0.033	-0.274	-0.283

Table 3-3 presents the sample from unitary matrices **U**.

Table 3-4: diagonal matrix.

0	**0**	**0**	**0**	**0**	**255.4872**
0	**0**	**0**	**0**	**105.5309**	**0**
0	**0**	**0**	**88.01752**	**0**	**0**
0	**0**	**80.92674**	**0**	**0**	**0**
0	**76.1892**	**0**	**0**	**0**	**0**
70.58321	**0**	**0**	**0**	**0**	**0**

Table 3-4: presents the sample from diagonal matrix **.**

Table 3.5: V^T the conjugate transpose of V(sample)

-0.017	0.018	-0.056	-0.02	0	-0.058
-0.031	-0.033	-0.059	-0.008	-0.006	-0.024
-0.294	-0.061	-0.185	0.038	-0.59	-0.218
0.047	-0.043	-0.096	0.028	-0.239	-0.125
-0.071	-0.038	0.009	0.058	-0.011	-0.037
0.053	-0.034	-0.025	-0.022	0.018	-0.04
0	0	0	0	0	0
0.016	-0.012	0.001	-0.021	0.011	-0.03
-0.014	-0.016	-0.034	-0.028	0.028	-0.049
-0.013	-0.046	0.052	0.077	0.052	-0.053
0	0.002	-0.005	-0.015	0.011	-0.022
0.001	0	0.002	-0.003	0.001	-0.001
0.006	0.129	-0.045	0.008	0.03	-0.018
-0.078	-0.037	0.031	0.073	0.047	-0.026
-0.013	-0.008	0.005	0.014	0.007	-0.002
-0.001	-0.001	0	-0.004	-0.002	-0.003
0.065	-0.016	0.119	-0.153	0.082	-0.056
0.011	-0.007	0.331	-0.427	0.203	-0.115
0.025	-0.016	0.01	0.023	0.009	-0.01
0.002	0.002	0.008	-0.006	0.006	-0.003
-0.027	0.071	-0.003	0.046	0.002	-0.101
0.054	-0.044	0.014	0.044	0.016	-0.144
0.001	0.032	-0.019	-0.03	0.035	-0.035

Table 3-5 presents the sample from conjugate transpose $\mathbf{V^T}$

Table 3-6: The original matrix M'

concepts Cvs	concept.1	concept.2	concept.3	concept.4	concept.5	concept.6	concept.7		concept.n
cv1	5.16	0.27	-1.41	-0.13	0.03	0.32	0.03	* * * * * * * *	1.46
cv2	1.85	1.61	0.44	-2.03	8.4	0.83	0.02	* * * * * * * *	0.99
cv3	1	0.44	3.62	2.25	0.48	0.71	0	* * * * * * * *	0.35
cv4	2.71	0.1	-0.48	2.1	0.81	2.64	0.02	* * * * * * * *	1.63
cv5	0.43	0.06	-0.27	0.33	0.17	0.39	0	* * * * * * * *	0.22
cv6	2.31	0.01	1.9	3.12	1.35	2.09	0.02	* * * * * * * *	1.28
cv7	1.56	0.31	-1.21	0.07	0.34	0.66	0.01	* * * * * * * *	0.57
cv8	2.6	2.67	0.37	0.53	0.04	2.35	0.01	* * * * * * * *	0.95
cv9	1.18	0.49	-0.04	0.43	0.56	0.74	0.01	* * * * * * * *	0.47
cv10	2.1	0.21	14.4	0.82	1.79	-0.36	0.01	* * * * * * * *	1.3
*	0.95	0.26	-1.07	0.07	0.33	0.62	0.01	* * * * * * * *	0.4
*	2.37	2.02	0.15	0.65	0.03	1.92	0.01	* * * * * * * *	0.85
*	1.14	0.46	4.92	0.73	0.83	0.13	0.01	* * * * * * * *	0.48
j.p1	2.17	1.12	13.5	8.39	0.86	1.93	0	* * * * * * * *	0.67
j.p2	0.9	0.46	4.7	2.38	0.73	0.54	0	* * * * * * * *	0.29
*	*	*	*	*	*	*	*	* * * * * * * *	*
j.pn	1.67	0.74	7.99	6.4	0.7	1.95	0.01	* * * * * * * *	0.62

Table 3-6 presents the final matrix of vectors M' that will be used in calculating the similarity between CVs and job posting.

In this matrix we see the purpose of applying SVD which is to use the reduced dimensionality representation to identify similar concepts and CVs. From matrix M' each CV represents as a vector (row in M') .

3.3 Finding Similarity To Match CVs And Job Posting.

In the previous step, by applying the LSA and SVD, we produced a vector for each C.V and for each job posting. Each of these vectors represents a semantic description of the corresponding C.V and job posting. These vectors are used in finding the similarity between CVs job postings, by applying the "Cosine Similarity" on between two vectors, one representing the job posting, and one representing the CV.

Applying the "Cosine Similarity" results in a value between 0 and 1, with number 0 denoting that there is no similarity and number 1 denoting that two vectors are identical. The equation for applying the Cosine Similarity is shown in Figure 3-7.

$$\text{Cosine}\ (v1,v2) = \frac{V1 * V\ 2}{\|V\ 1\| * \|V2\|}.$$

Figure 3-7 :Equation of cosine similarity.

Where V1: CV vector, V2: job posting vector .

Chapter four

Clustering of CVs

This chapter briefly presents the background of the clustering technique and presents how our approach enhanced the clustering process by building the clusters that are based on the vectors in the steps outlined in section 3.3. What distinguishes our enhanced clustering methodology is that the vectors are based on concepts and instances.

4.1Clustering

Clustering is considered to be one of the most important techniques in data mining. The algorithms and techniques of clustering's enables to separate data items and objects into groups called clusters. Members of each cluster, have similar properties, (Konda .S 2007), Bandyopadhyay, S.(2002).

Clustering is also defined as a statistical technique used to build classes as one structure with each class having a set of details that have a high degree of association between members of the same class, (Anderberg, M.R.(1973)). Figure 4.1 :shows an example of Clustering :

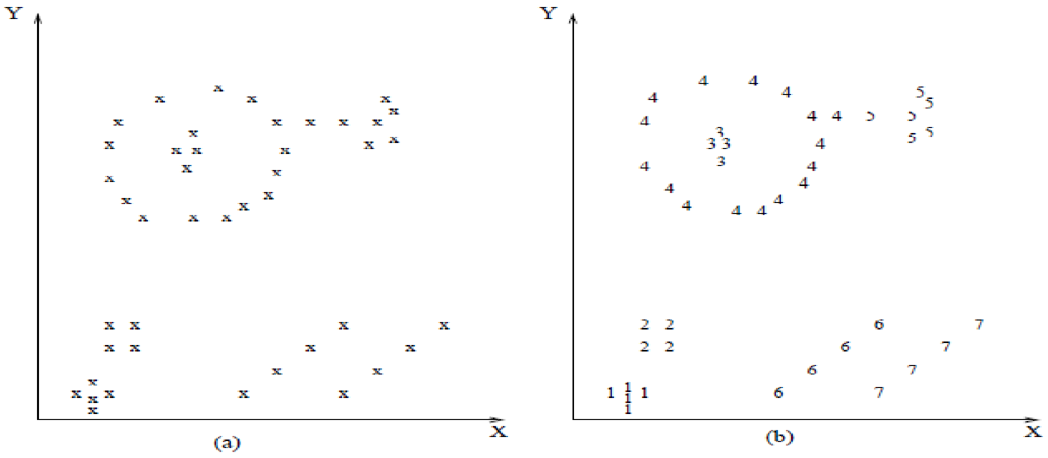


Figure4-1:DataClustering.

According to Berkhin, P. (2006), there are two main types or techniques of clustering:

1- The simple non-hierarchical methods (partitioning): This method divides the collection of text that contains N objects into M clusters. In this method meta overlapping isn't allowed and each object membership is assigned to only one cluster, which is the closest group or cluster.

2- Hierarchical methods : This method produces nested datasets like a tree, where each pairs of documents or clusters are successively linked together until every cluster in the data set is connected.

The basic steps to develop clustering process according to Halkidi, M.,et al(2001). are presented in figure5-2.These steps can be summarized as follows:

1- Document representation: The goal of this step is to properly represent the features in documents and files in a structure that is acceptable by the clustering process.

2- The clustering algorithm: The goal of this step refers to choose the most appropriate clustering algorithms. Examples of existing clustering algorithms are; K-Means, Fuzzy.

4.2 Clustering of CVs and job posting :

Clustering in the field of e-recruitment aids in the filtration process, by dividing the CVs and job postings into clusters. Each cluster contains similar CVs and job postings.

A typical document clustering activity involves the following steps(Jain and Dubes1988):

(1) Document representation (optionally including feature extraction and/or selection).

(2) Definition of a document proximity measure appropriate to the data domain.

(3) Clustering or grouping (clustering algorithm selection).

(4) Data abstraction (if needed).

(5) Validation of results.

Figure 4-2 : Show the basic steps of the clustering process

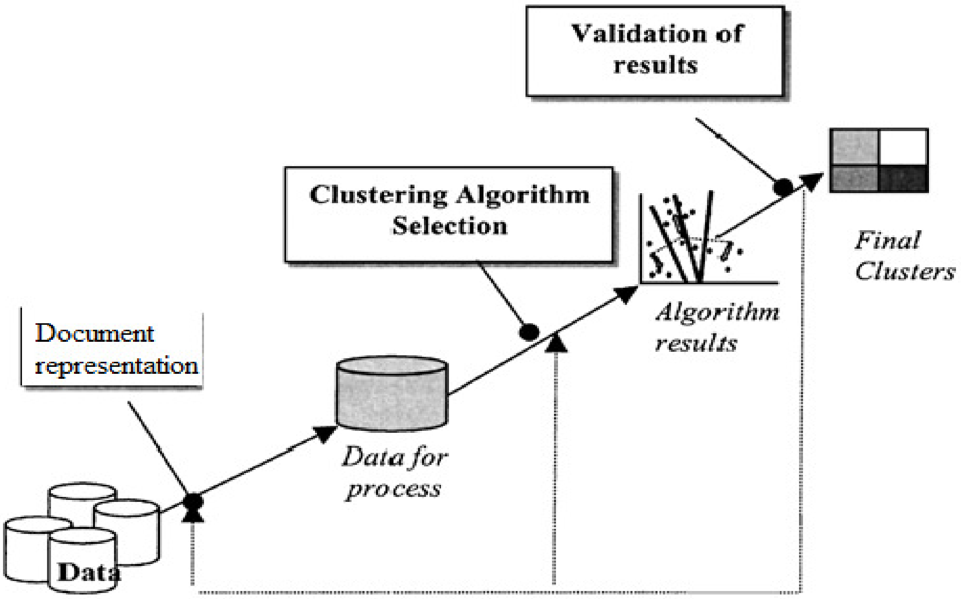
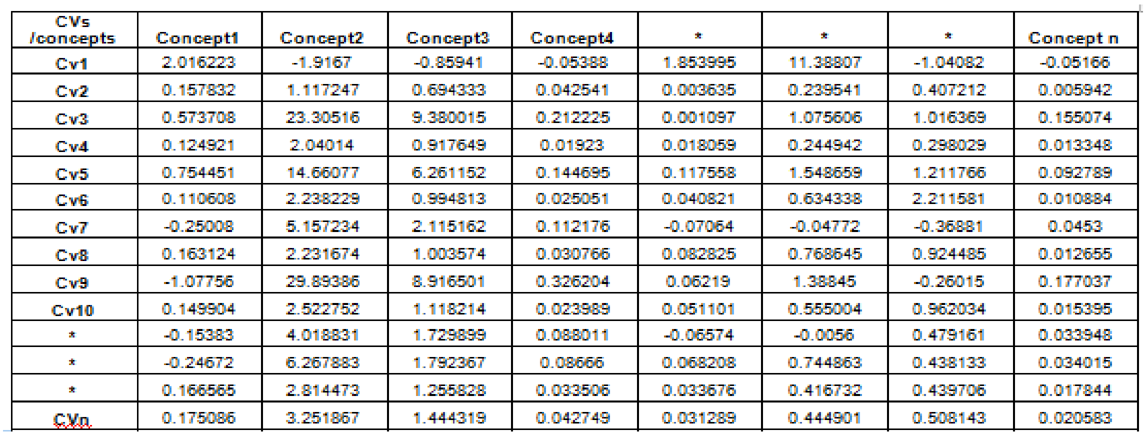


Figure 4-2 :clustering process.

Our research enhanced the traditional clustering process by providing the vectors resulting from the steps outlined in section 3.3 (which are more accurate because they are based on concepts and instances) as inputs to the clustering process. Table 4-1 shows fragments of the vectors of CVs and job postings are passed as inputs to the clustering process.

Table 4-1 : The vectors of CVs.



CVs /concepts	Concept1	Concept2	Concept3	Concept4	*	*	*	Concept n
Cv1	2.016223	-1.9167	-0.85941	-0.05388	1.853995	11.38807	-1.04082	-0.05166
Cv2	0.157832	1.117247	0.694333	0.042541	0.003635	0.239541	0.407212	0.005942
Cv3	0.573708	23.30516	9.380015	0.212225	0.001097	1.075606	1.016369	0.155074
Cv4	0.124921	2.04014	0.917649	0.01923	0.018059	0.244942	0.298029	0.013348
Cv5	0.754451	14.66077	6.261152	0.144695	0.117558	1.548659	1.211766	0.092789
Cv6	0.110608	2.238229	0.994813	0.025051	0.040821	0.634338	2.211581	0.010884
Cv7	-0.25008	5.157234	2.115162	0.112176	-0.07064	-0.04772	-0.36881	0.0453
Cv8	0.163124	2.231674	1.003574	0.030766	0.082825	0.768645	0.924485	0.012655
Cv9	-1.07756	29.89386	8.916501	0.326204	0.06219	1.38845	-0.26015	0.177037
Cv10	0.149904	2.522752	1.118214	0.023989	0.051101	0.555004	0.962034	0.015395
*	-0.15383	4.018831	1.729899	0.088011	-0.06574	-0.0056	0.479161	0.033948
*	-0.24672	6.267883	1.792367	0.08666	0.068208	0.744863	0.438133	0.034015
*	0.166565	2.814473	1.255828	0.033506	0.033676	0.416732	0.439706	0.017844
CVn	0.175086	3.251867	1.444319	0.042749	0.031289	0.444901	0.508143	0.020583

4.3 Experiment result

To investigate the validity of our enhanced clustering approach, we applied our clustering technique, in which vectors(the inputs to the clustering algorithm) are based on concepts and instances, on a set of CVs. We also applied the traditional clustering technique in which CV vectors are based on keywords. The results of the two applications are compared to the clustering of a human expert. We use K-Means as the clustering algorithm.

Table 4-2 : Shows the clustering by the human expert. This table is used to judge the accuracy of both clustering methods under question. The next sub-sections in this section show the results of the two clustering methods.

Table 4-2: Expert clustering of CVs and job posting.

cluster	matching CVs	Total matches
(1)	J.p1,CV12,CV17, CV13,CV14	5
(2)	J.p2,CV1,CV6,CV8,CV10	5
(3)	J.p3,CV3, CV2,CV16,CV18,CV19,CV20	7
(4)	j.p4,CV4,CV5,CV9,CV15	5
(5)	CV7,CV11, J.p5	3

4.3.1 Clustring based on Concepts and Instances.

Table 4-3 presents the results of clustering 20 CVs and 5 job postings in our approach.

The K-means algorithm was implanted using an open source program written in C# language. We had to make small modifications to the code to implement our approach.

Table 4-3: Clustring based on Concepts and Instances.

Clusters		1	2	3	4	5
***CVs* and Job posting in cluster**	CVs	CV13 CV14 CV17 CV12	CV3 CV8 CV15 CV1 CV6	CV19 CV2 CV16 CV20	CV4 CV9 CV18 CV5	CV7 CV11 CV10
	j.ps	j.p1	j.p2	j.p3	j.p4	j.p5

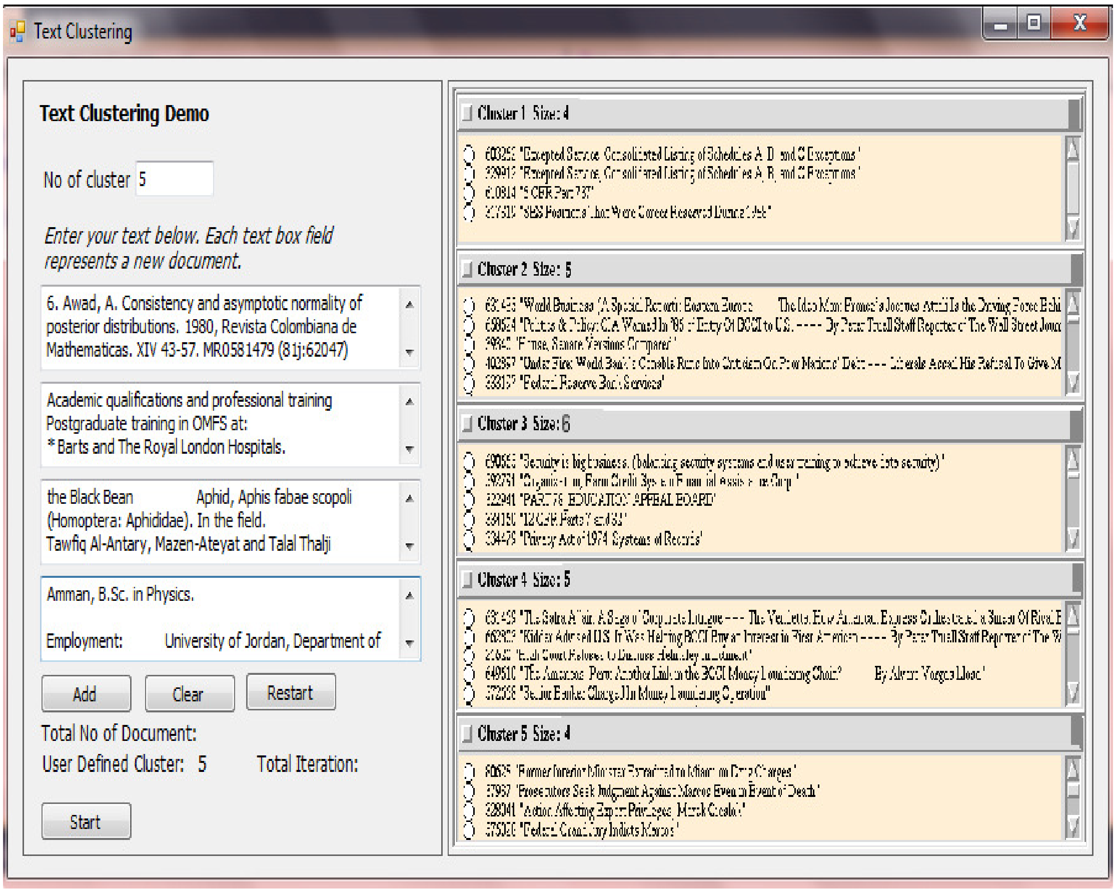
In table 4-3, we can clearly see that each cluster have many CVs and one job posting . Each job posting has good similarity with many CVs.

Compared to the experts clustering, 20 out of 25 CVs and job postings was put to its correct cluster, which means a success rate of 80 %.

4.3.2 Clustring by Tradational Approach.

Table 4-4: shows the results of clustering 20 CVs and 5 job postings according to the traditional keyword approach. Those CVs were the same used in section 4.2.

(used TextClustering tool). We used Text Clustering as cluster tool and K-Means as the clustering algorithm. Figure 4-3:shows interface for TextClustering tool.



No of cluster 5

Enter your text below. Each text box field represents a new document.

6. Awad, A. Consistency and asymptotic normality of posterior distributions. 1980, Revista Colombiana de Mathematicas. XIV 43-57. MR0581479 (81j:62047)

Academic qualifications and professional training
Postgraduate training in OMFS at:
* Barts and The Royal London Hospitals.

the Black Bean Aphid, Aphis fabae scopoli (Homoptera: Aphididae). In the field.
Tawfiq Al-Antary, Mazen-Ateyat and Talal Thalji

Amman, B.Sc. in Physics.

Employment: University of Jordan, Department of

Total No of Document:

User Defined Cluster: 5 Total Iteration:

Cluster 1 Size: 4

○ 603252 "Excepted Service Consolidated Listing of Schedules A, B, and C Exceptions"
○ 229912 "Excepted Service, Consolidated Listing of Schedules A, B, and C Exceptions"
○ 610814 "5 CFR Part 737"
○ 207519 "SES Positions That Were Career Reserved During 1988"

Cluster 2 Size: 5

○ 631495 "World Business (A Special Report): Eastern Europe --- The Idea Man: France's Jacques Attali Is the Driving Force Behi
○ 658524 "Politics & Policy: CIA Warned In '85 of Entry Of BCCI to U.S. ---- By Peter Truell Staff Reporter of The Wall Street Jour
○ 29842 "House, Senate Versions Compared"
○ 402857 "Under Fire: World Bank's Conable Runs Into Criticism On Poor Nations' Debt --- Liberals Assail His Refusal To Give M
○ 333157 "Federal Reserve Bank Services"

Cluster 3 Size: 6

○ 690565 "Security is big business. (balancing security systems and user training to achieve data security)"
○ 392731 "Organization; Farm Credit System Financial Assistance Corp."
○ 322941 "PART 78 EDUCATION APPEAL BOARD"
○ 334150 "12 CFR Parts 7 and 32"
○ 334475 "Privacy Act of 1974 Systems of Records"

Cluster 4 Size: 5

○ 631459 "The Safra Affair: A Saga of Corporate Intrigue --- The Vendetta: How American Express Orchestrated a Smear Of Rival B
○ 662822 "Kidder Advised U.S. It Was Helping BCCI Buy an Interest in First American ---- By Peter Truell Staff Reporter of The W
○ 21632 "High Court Refuses to Dismiss Helmsley Indictment"
○ 649510 "The Americas: Peru: Another Link in the BCCI Money Laundering Chain? By Alvaro Vargas Llosa"
○ 372058 "Senior Banker Charged In Money Laundering Operation"

Cluster 5 Size: 4

○ 80625 "Former Interior Minister Extradited to Miami on Drug Charges"
○ 37937 "Prosecutors Seek Judgment Against Marcos Even in Event of Death"
○ 328041 "Action Affecting Export Privileges; Marek Cecala"
○ 375020 "Federal Grand Jury Indicts Marcos"

Figure 4-3: Text Clustering tool.

Table 4-4: show the result of clustering CVs in traditional approach.

Clutters		1	2	3	4	5
***CVs* and Job posting in cluster**	CVs	CV3 CV13 CV14	CV8 CV15 CV1 CV12	CV2 CV6 CV10 CV20 CV19	CV4 CV9 CV18 CV5	CV16 CV7 CV11
	j.ps	j.p 1	j.p2	j.p3	j.p4	j.p5

By comparing the result in table 4-3 to the experts judgment , we can see that 7 CVs were not put in right cluster and 18 CVs and job postings were correctly clustered. The success rate is 72% .

Chapter five

Experimental Results

To investigate the validity the CVMM, we applied it, on a set of CVs. We also applied the other traditional approaches(shown in table 5-1) on the same set. The size of the sample was 44 CVs. The results of these applications are compared to the judgment of a human expert.

Measure	Description
Expert	The similarity measure obtained from experts judgment (as a base measure).
Keyword	similarity measure between common keywords in job posting and CVs (without LSA)
COS-Word **(Traditional LSA)**	Cosine similarity measure between CVs and job posting using all words as a vector (with LSA)
COS-(Concepts only)	Cosine similarity measure between CVs and job posting using concepts only as a vector (with LSA)
CVMM	Cosine similarity measure between CVs and job posting using concepts and instances (with LSA)

Table 5-1: Glossary of approaches.

This chapter is organized into sections with each section showing the similarity between samples of each the CVs and the matched job posting in each approach. It also compares the result of applying the approach with the result of the expert judgment.

5.1 The expert judgment (keyword):

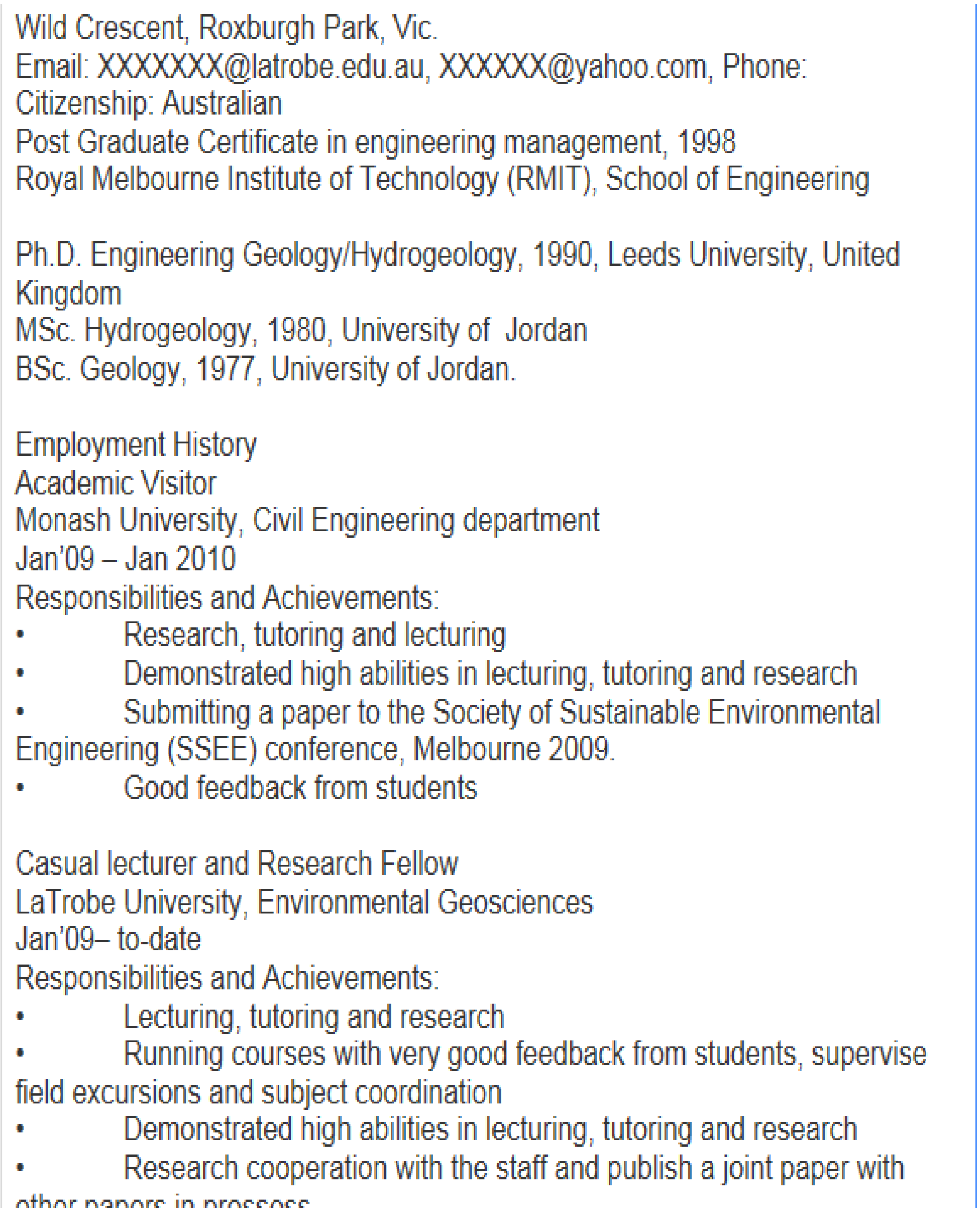
In order to correctly judge the accuracy of our approaches, we have to compare them with the judgment and results of an expert. The overall number of CVs was 44 and the overall number of job postings was 5. The job postings were related to the following field : Engineering, Mathematics, Animals and Plants, Chemistry, and Physics.

Table 5-2 presents the job postings in the conducted experiments along with CVs that are classified to be a match for each job posting. Each row, represents all the CVs that match one job posting. For example, the CVs (CV7,CV11, CV44) are classified to match the job posting(J.p5).

Table 5-2 :Expert CVs-Job postings matching.

Job posting	matching CVs	**Total matches**
Engineering (J.p1)	CV12,CV17,CV22,CV24,CV13,CV14,CV37	7
Mathematical (J.p2)	CV1,CV6,CV8,CV10,CV26,CV27,CV40,CV42,CV25	9
Animal and plant (J.p3)	CV3, CV2,CV16,CV18,CV19,CV20,CV21,CV23 ,CV30,CV31,CV32,CV33,CV34,CV35,CV36	15
Chemistry (J.p4)	CV4 ,CV5,CV9,CV15,CV28,CV29,CV38,CV39,CV41, CV43	10
Physics (J.p5)	CV7,CV11, CV44	3

Figure 5-1 :shows the CV12 that are matching the engineering job posting



Wild Crescent, Roxburgh Park, Vic.
Email: XXXXXXX@latrobe.edu.au, XXXXXX@yahoo.com, Phone:
Citizenship: Australian
Post Graduate Certificate in engineering management, 1998
Royal Melbourne Institute of Technology (RMIT), School of Engineering

Ph.D. Engineering Geology/Hydrogeology, 1990, Leeds University, United Kingdom
MSc. Hydrogeology, 1980, University of Jordan
BSc. Geology, 1977, University of Jordan.

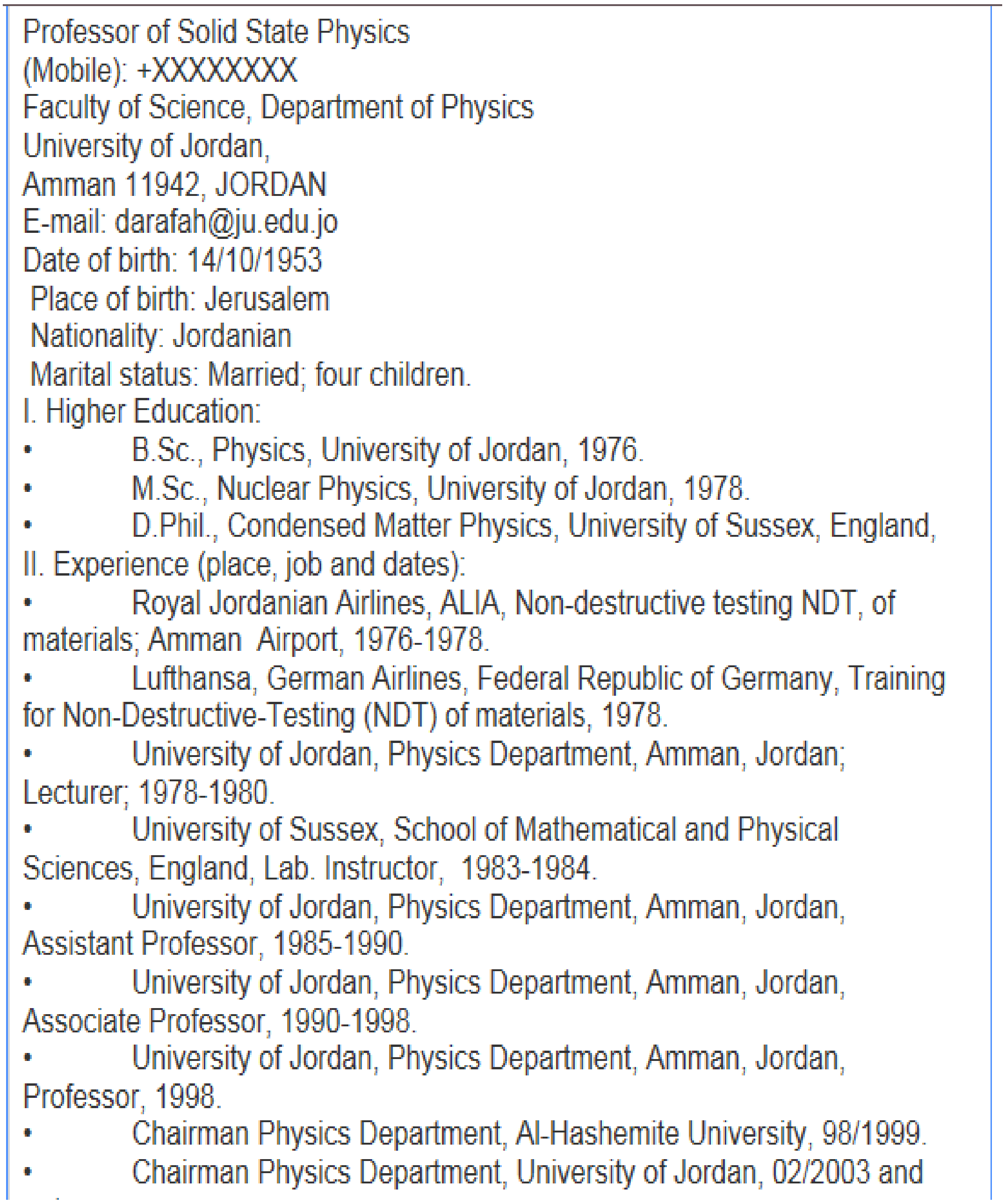
Employment History
Academic Visitor
Monash University, Civil Engineering department
Jan'09 – Jan 2010
Responsibilities and Achievements:

- Research, tutoring and lecturing
- Demonstrated high abilities in lecturing, tutoring and research
- Submitting a paper to the Society of Sustainable Environmental Engineering (SSEE) conference, Melbourne 2009.
- Good feedback from students

Casual lecturer and Research Fellow
LaTrobe University, Environmental Geosciences
Jan'09– to-date
Responsibilities and Achievements:

- Lecturing, tutoring and research
- Running courses with very good feedback from students, supervise field excursions and subject coordination
- Demonstrated high abilities in lecturing, tutoring and research
- Research cooperation with the staff and publish a joint paper with other papers in process

We see from figure 5-1, CV12 Belonging to a professor of engineering.



Professor of Solid State Physics
(Mobile): +XXXXXXXX
Faculty of Science, Department of Physics
University of Jordan,
Amman 11942, JORDAN
E-mail: darafah@ju.edu.jo
Date of birth: 14/10/1953
Place of birth: Jerusalem
Nationality: Jordanian
Marital status: Married; four children.

I. Higher Education:

- B.Sc., Physics, University of Jordan, 1976.
- M.Sc., Nuclear Physics, University of Jordan, 1978.
- D.Phil., Condensed Matter Physics, University of Sussex, England,

II. Experience (place, job and dates):

- Royal Jordanian Airlines, ALIA, Non-destructive testing NDT, of materials; Amman Airport, 1976-1978.
- Lufthansa, German Airlines, Federal Republic of Germany, Training for Non-Destructive-Testing (NDT) of materials, 1978.
- University of Jordan, Physics Department, Amman, Jordan; Lecturer; 1978-1980.
- University of Sussex, School of Mathematical and Physical Sciences, England, Lab. Instructor, 1983-1984.
- University of Jordan, Physics Department, Amman, Jordan, Assistant Professor, 1985-1990.
- University of Jordan, Physics Department, Amman, Jordan, Associate Professor, 1990-1998.
- University of Jordan, Physics Department, Amman, Jordan, Professor, 1998.
- Chairman Physics Department, Al-Hashemite University, 98/1999.
- Chairman Physics Department, University of Jordan, 02/2003 and

Figure 5-2 : CV7 that is matching physics job posting:

5.2 The similarity (keyword) :

Table 5-3: shows fragment of the results of matching according to the keyword approach.

CVs	J.Ps	similarity	Expert	CVs	J.ps	similarity	Expert
Cv1	**J.P1**	0.83	No	CV11	**J.P1**	0.77	No
Cv2	**J.P3**	0.84	Yes	CV12	**J.P1**	0.74	Yes
Cv3	**J.P1**	0.72	No	CV13	**J.P1**	0.79	Yes
Cv4	**J.P1**	0.81	No	CV14	**J.P1**	0.74	Yes
CV5	**J.P1**	0.87	No	CV15	**J.P3**	0.83	No
CV6	**J.P3**	0.68	No	CV16	**J.P3**	0.80	Yes
CV7	**J.P1**	0.68	No	CV17	**J.P3**	0.77	No
CV8	**J.P2**	0.83	Yes	CV18	**J.P1**	0.70	No
CV9	**J.P1**	0.77	No	CV19	**J.P3**	0.83	Yes
CV10	**J.P3**	0.72	No	CV20	**J.P3**	0.82	Yes

In this section we used all keywords in CVs and job posting to find the similarity. We extracted all the words from each job posting after removing stopping words and comparing and calculating how many word from these words appear in each CV (Ex: the first job posting contain around 50 word when match with first CV just 16 word appear). To find the similarity between CV and job posting we compute the percentage of words appearing in the CV to the total number of words in the job posting.

5.3 COS-Word (Traditional LSA) :

Table 5-4: shows fragment of the results of matching according to the LSA(Word) approach.

CVs	J.Ps	similarity	Expert	CVs	J.ps	similarity	Expert
Cv1	J.P5	0.93	No	CV11	J.P3	0.93	No
Cv2	J.P3	0.71	Yes	CV12	J.P3	0.93	No
Cv3	J.P3	0.84	Yes	CV13	J.P1	0.87	yes
Cv4	J.P3	0.64	No	CV14	J.P4	0.87	No
CV5	J.P3	0.85	No	CV15	J.P3	0.75	No
CV6	J.P2	0.87	Yes	CV16	J.P4	0.75	No
CV7	J.P3	0.69	No	CV17	J.P4	0.48	No
CV8	J.P4	0.62	No	CV18	J.P3	0.71	Yes
CV9	J.P4	0.71	Yes	CV19	J.P3	0.71	Yes
CV10	J.P3	0.88	No	CV20	J.P3	0.94	Yes

5.4 COS-(Concepts only):

Table 5-5 presents a fragment from result of the matching in the LSA with Concepts approach (without instances).

CVs	J.Ps	similarity	Expert	CVs	J.ps	similarity	Expert
Cv1	J.P2	0.89	Yes	CV11	J.P1	0.79	No
Cv2	J.P3	0.77	Yes	CV12	J.P5	0.83	No
Cv3	J.P4	0.86	No	CV13	J.P1	0.89	No
Cv4	J.P5	0.87	No	CV14	J.P1	0.87	No
CV5	J.P5	0.88	No	CV15	J.P1	0.85	Yes
CV6	J.P3	0.89	No	CV16	J.P4	0.85	Yes
CV7	J.P1	0.87	No	CV17	J.P4	0.85	No
CV8	J.P1	0.90	No	CV18	J.P4	0.74	No
CV9	J.P5	0.70	No	CV19	J.P4	0.86	No
CV10	J.P3	0.90	No	CV20	J.P4	0.83	No

5.5 CVMM (Our proposed approach):

Table 5-6 presents a fragment from result of the matching in our approach (Ontology that is based on concepts and instances).

Table 5-6: the result of matching using LSA and ontology.

CVs	J.Ps	similarity	Expert	CVs	J.ps	similarity	Expert
Cv1	J.P2	0.96	Yes	CV11	J.P5	0.78	Yes
Cv2	J.P3	0.75	Yes	CV12	J.P1	0.84	Yes
Cv3	J.P3	0.94	Yes	CV13	J.P1	0.82	Yes
Cv4	J.P5	0.87	No	CV14	J.P3	0.84	No
CV5	J.P4	0.76	Yes	CV15	J.P3	0.78	No
CV6	J.P2	0.94	Yes	CV16	J.P3	0.77	No
CV7	J.P5	0.74	Yes	CV17	J.P1	0.86	Yes
CV8	J.P5	0.85	No	CV18	J.P4	0.90	Yes
CV9	J.P1	0.82	No	CV19	J.P4	0.91	Yes
CV10	J.P2	0.85	Yes	CV20	J.P4	0.92	Yes

5.6 Analysis of experimental results.

This section summarizes and analyzes the results presented in this chapter.

Table 5-7: shows the success rates of all other approaches.

Approaches	**Success Rate**
Keyword	**38%**
COS-Word **(Traditional LSA)**	**52%**
COS-(Concepts only)	**25%**
CVMM	**84%**

The first approach(keywords) has classified 17 CVs correctly and 27 CVs incorrectly. The accuracy rate was 38% and the mismatch rate was 62%.

The second approach(traditional LSA) has classified 23 CVs correctly and 21 CVs incorrectly. The accuracy rate was 52% and the mismatch rate was 48%.

The third approach(concepts only) has classified 11 CVs correctly and 33 CVs incorrectly. The accuracy rate was 25% and the mismatch rate was 75%. These results reveal that using ontology based on concepts did not enhance the matching process, nor it enhanced the LSA. The reason of this is that many of concepts extracted from the CVs and job postings were eliminated in the ontology building process because of its low frequency.

Our approach has classified 37 CVs correctly and 7 CVs incorrectly. The accuracy rate was 84% and mismatch rate was 16%. It is clear from these results that our approach achieved the highest accuracy rate. We attribute this high accuracy to the usage of ontology that is based on instances, which is the core idea of this thesis.

Figure 5-3: shows a chart for result of all approaches.

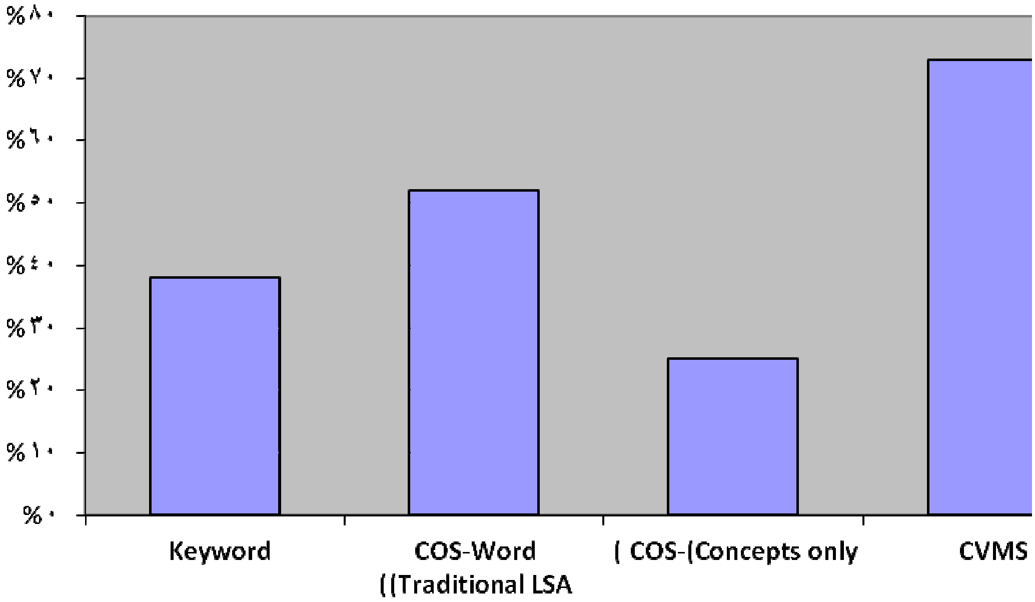


Figure 5-3: Chart for result of all approaches.

Chapter Six

Conclusion and Future Work

6.1 Conclusion

The proposed CV matching methodology could match between CVs and job posting depending on their semantic similarity and the process of matching is able to match between CVs and job postings from all domains. The proposed methodology does not impose a specific form to write the CV and the job posting. Our matching methodology proposes two techniques for matching, one dependent on the cosine similarity, and the other dependent on the clustering to calculate the distance between a CV and a job posting.

Combining ontological concepts and LSA, as performed by our methodology, produced more accurate results than all other traditional approaches. We were able to prove the accuracy and efficiency of our approach by applying it on a sample of CVs and job postings and calculating the similarity, and comparing the results from applying other approaches on the same sample. The success rates of the proposed matching system were the highest among all conducted experiments. The traditional approaches in which we compared our work are : the keyword matching approach, the traditional LSA approach and the LSA by concept approach.

6.2 Recommendations and Future work

This research focused on the e-recruitment domain. Covering more domains using the proposed system, will enable more and more domains to utilize our methodology for better efficiency. Also, achieving more accurate results is still a topic of continuous and constant research.

References

Amdouni S., and Karaa W. (2010, May). Web-based recruiting. In *Computer Systems and Applications (AICCSA), 2010 IEEE/ACS International Conference on* (pp. 1-7). IEEE.

Anderberg, M. R. (1973). Cluster analysis for applications (No. OAS-TR-73-9). OFFICE OF THE ASSISTANT FOR STUDY SUPPORT KIRTLAND AFB N MEX.

Bandyopadhyay, S., & Maulik, U. (2002). An evolutionary technique based on K-means algorithm for optimal clustering in RN. Information Sciences, 146(1), 221-237.

Baraldi, A., & Alpaydin, E. (2002). Constructive feedforward ART clustering networks. I. Neural Networks, IEEE Transactions on, 13(3), 645-661.

Berkhin, P. (2006). A survey of clustering data mining techniques. In Grouping multidimensional data (pp. 25-71). Springer Berlin Heidelberg.

Bettahar, F., Moulin, C., & Barthès, J. P. (2009). Towards a Semantic Interoperability in an E-government Application. Electronic Journal of E-government, 7(3), 209-226.

Berry, M. W., Dumais, S. T., & O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. SIAM review, 37(4), 573-595.

Bizer C., Heese, R., Mochol, M., Oldakowski, R., Tolksdorf, R., & Eckstein, R. (2005). The impact of semantic web technologies on job recruitment processes. *Wirtschaftsinformatik 2005*, 1367-1381.

Cosma, G., & Joy, M. (2012). An approach to source-code plagiarism detection and investigation using latent semantic analysis. Computers, IEEE Transactions on, 61(3), 379-394.

Cosma, G., & Joy, M. S. (2012). Evaluating the Effectiveness Latent Semantic Analysis for Similar Source-code Detection. Informatica.

Davies, P. (1971). New views of lexicon. In J. B. Carroll, P. Davies, &B. Richman (Eds.), Word frequency book (pp. xli-liv). New York : Houghton Mifflin and American Heritage.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American society for information science, 41(6), 391-407.

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., & Harshman, R. (1988, May). Using latent semantic analysis to improve access to textual information. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 281-285). ACM.

East E. William (2008)IMPACT OF E-GOVERNMENT ON FEDERAL FACILITY DELIVERY [Conference]. - Santiago, Chile : [s.n.],.

Fazel-Zarandi, M.; Fox, M.S, (2009), "Semantic Matchmaking for Job Recruitment: An Ontology-Based Hybrid Approach", in Proceedings of the 3rd *International Journal of Computer Applications (0975 – 8887) Volume 51– No.2, August 2012* 45.

Fonou-Dombeu, J. V., & Huisman, M. (2011). Semantic-Driven e-Government: Application of Uschold and King Ontology Building Methodology for Semantic Ontology Models Development. arXiv preprint arXiv:1111.1941.

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1984, May). Statistical semantics: Analysis of the potential performance of keyword information systems. In Human factors in computer systems (pp. 187-242). Ablex Publishing Corp.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. Knowledge acquisition, 5(2), 199-220.

Hassan F. M., Ghani, I., Faheem, M., & Hajji, A. A. (2012). Ontology Matching Approaches for eRecruitment. *International Journal of Computer Applications*, *51*(2), 39-45.

Hexin, L., & Bin, Z. (2010). Elastic information matching technology and its application in electronic recruitment. In *Computer-Aided Industrial Design & Conceptual Design (CAIDCD), 2010 IEEE 11th International Conference on* (Vol. 2, pp. 1582-1585). IEEE.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. Machine Learning, 42(1), 177-196.

Huang, A. (2008). Similarity measures for text document clustering. Proceedings of NZCSRSC, 49-56.

Hu, X., Zhang, X., Lu, C., Park, E. K., & Zhou, X. (2009, June). Exploiting Wikipedia as external knowledge for document clustering. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 389-396). ACM.

Kayed, A., Hirzalla, N., Ahmad, H., & Al Faisal, E. (2011). Extracting Concepts for Software Components. Trends in Network and Communications, 694-699.

Kayed A., El-Qawasmeh, E., & Qawaqneh, Z. (2010). Ranking web sites using domain ontology concepts. Information & management, 47(7), 350-355.

Kayed, A., Nizar, M., & Alfayoumi, M. (2010, May). Ontology concepts for requirements engineering process in e-government applications. In Internet and Web Applications and Services (ICIW), 2010 Fifth International Conference on (pp. 396-400). IEEE.

Keim, T., König, W., von Westarp, F., Weitzel, T., & Wendt, O. (2005). Recruiting Trends 2005-Eine empirische Untersuchung der Top-1.000-Unternehmen in Deutschland und von 1.000 Unternehmen aus dem Mittelstand. Working paper.

Landauer, T.K., Dumais, S.T. (1997): A Solution to Plato's Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge. Psychological Review,104 , 211-240.

Liao, C. H., Kuo, B. C., & Pai, K. C. (2012). effectiveness of automated chinese sentence scoring with latent semantic analysis. tojet, 11(2).

Lintean, M., Moldovan, C., Rus, V., & McNamara, D. (2010, June). The role of local and global weighting in assessing the semantic similarity of texts using Latent Semantic Analysis. In Proceedings of the 23rd International Florida Artificial Intelligence Research Society Conference. Daytona Beach, FL.

Lintean, M., Moldovan, C., Rus, V., & McNamara, D. (2010, June). The role of local and global weighting in assessing the semantic similarity of texts using Latent Semantic Analysis. In Proceedings of the 23rd International Florida Artificial Intelligence Research Society Conference. Daytona Beach, FL.

López, M. F., Gómez-Pérez, A., Sierra, J. P., & Sierra, A. P. (1999). Building a chemical ontology using methontology and the ontology design environment. Intelligent Systems and their Applications, IEEE, 14(1), 37-46.

Lv, H., & Zhu, B. (2006, November). Skill ontology-based semantic model and its matching algorithm. In Computer-Aided Industrial Design and Conceptual Design, 2006. CAIDCD'06. 7th International Conference on (pp. 1-4). IEEE.

Marksberry, P., Church, J., & Schmidt, M. (2012). The employee suggestion system: A new approach using latent semantic analysis. Human Factors and Ergonomics in Manufacturing & Service Industries.

Mochol, M.; Jentzsch, A. Wache, H, (2007), “ Suitable employees wanted? Find them with semantic techniques”, In Proceedings of Workshop on Making Semantics Web For Business at European Semantic Technology Conference (ESTC2007)", Vienna, Austria.

Mochol, M., Oldakowski, R., Heese, R., & und Informationssysteme, D. (2004, September). Ontology-based Recruitment Process. In Proc. of the GI2004Conference,Ulm,Germany.http://page.mi.fuberlin.de/mochol/papers/SemTech.pd.

Mount, D. (2005). KMlocal: a testbed for k-means clustering algorithms.

Nanopoulos, A., Theodoridis, Y., & Manolopoulos, Y. (2001, September). C2P: Clustering based on closest pairs. In PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES (pp. 331-340).

Pandey, A. (2006). *Staffing Management*. Global Vision Publishing Ho. ,isbn:81-8220-179-9.

Salhofer, P., Stadlhofer, B., & Tretter, G. (2009, September). Ontology driven e-government. In Software Engineering Advances, 2009. ICSEA'09. Fourth International Conference on (pp. 378-383). IEEE.

Simperl, E. (2009). Reusing ontologies on the Semantic Web: A feasibility study. Data & Knowledge Engineering, 68(10), 905-925.

Tata, S., & Patel, J. M. (2007). Estimating the selectivity of tf-idf based cosine similarity predicates. ACM SIGMOD Record, 36(2), 7-12.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. Neural Networks, IEEE Transactions on, 16(3), 645-678.

Yahiaoui L., et al, (2006), “Semantic Annotation of Documents Applied to e-recruitment”, Proceedings of SWAP the 3rd Italian Semantic Web Workshop, Publisher: Citeseer, Pages: 1-6.

WEBSITES REFERENCES

http://www.mathworks.com/products/matlab/

http://www.sovren.com

http://ontogen.ijs.si/

http://protege.stanford.edu/overview/

http://wordnet.princeton.edu/

http://www.avidian.com/crm/crm-system-for-outlook.aspx

APPENDICES

Appendix A:

C++ Program to find frequency of concept and return a common concepts in CVs and job posting.

```
#include <iostream>
 #include <map>
 #include <set>
 #include <string>
 #include <fstream>
 #include "vector"
#include <algorithm>
 #include <string>
using namespace std;
int WordFreq(char *pcFileName,char  *pcFileName1,char *pcFileName2 ,char
*pcFileName3,int a) {
      vector<string> AllWordVec; // vector with all word inc duplicates
     set<string> UniqueWords; // set with only unique words
    map<string, int> WordFreq; // Map of words and their frequencies
    string word; // Used to hold input word
           if (0 == pcFileName)  {
cout << "Null input provided" << endl;
                   return 1; }
// Open the input text file
 ifstream InputFile(pcFileName);
ifstream InputFile1(pcFileName1);
ofstream OutputFile2(pcFileName2);
ofstream OutputFile3(pcFileName3);
if (!InputFile.is_open()) {
 cout << "Can not open the file" << endl;
 return 1;
 }
if (!InputFile1.is_open()) {
 cout << "Can not open the file" << endl;
 return 1;
 }
// Read all tokens from the input file
 while (InputFile >> word) {
 //-- Read unique words/tokens; this is set, so no duplicates
      std::string data = word ;
      std::transform(data.begin(), data.end(), data.begin(),::tolower );
     UniqueWords.insert(data);
 }
//-- Read all words/tokens in the sequence
 while (InputFile1 >> word)
 {
      std::string data = word ;
      std::transform(data.begin(), data.end(), data.begin(),::tolower );
      AllWordVec.push_back(data); }
vector<string>::iterator vWord; // for iterating thru all words
 set<string>::const_iterator UWord; // for iterating thru unique words
for (UWord = UniqueWords.begin(); UWord != UniqueWords.end(); ++UWord){
 // check for unique word in the AllWordVec vector
 for(vWord = AllWordVec.begin(); vWord != AllWordVec.end(); vWord++ )
 {
 if (!(*UWord).compare(*vWord))
 {
 // word matched, count this
 WordFreq[*UWord]++;
 }}}
// Write count, word
 map<string, int>::const_iterator iter;
 for (iter = WordFreq.begin(); iter != WordFreq.end(); ++iter) {
    a++;
OutputFile2<< iter->first <<<<  iter->second <<endl;
OutputFile3<<  iter->second << endl; }
            return a;}
// main function
 int main() {
 int q;
char*fnamame[50]={"1.txt","2.txt","3.txt","4.txt","5.txt","6.txt","7.txt","8.txt","9.txt","10.txt","11.txt","12.txt","13.txt","14.txt","15.txt","16.txt","17.txt","18.txt","19.txt","20.txt","21.txt","22.txt","23.txt","24.txt","25.txt","26.txt","27.txt","28.txt","29.txt","30.txt","31.txt","32.txt","33.txt","34.txt","35.txt","36.txt","37.txt","38.txt","39.txt","40.txt","41.txt","42.txt","43.txt","44.txt","45.txt","physics.txt","chemistry1.txt","anmailsciences.txt","math.txt","Engineering.txt"};
for (int i=0;i<45;i++)        {
cout <<endl<<fnamame [i]<<endl;
q=WordFreq("segaconcept.txt","segaconcept.txt","ee.txt","ee1.txt",0);
cout<<""<<endl;
//cout<<q;}
 system("pause");
 return 0; }
```

Appendix B :

1. the result of keyword approach (similarity between CVs and job posting).

Job.posting CVs	J.p1	J.p2	J.p3	J.p4	J.p5	expert
CV1	0.83	0.73	0.8	0.7	0.73	No
CV2	0.77	0.69	0.84	0.72	0.74	Yes
CV3	0.72	0.61	0.65	0.55	0.63	No
CV4	0.81	0.68	0.77	0.69	0.7	No
CV5	0.87	0.73	0.69	0.63	0.78	No
CV6	0.67	0.62	0.68	0.57	0.66	No
CV7	0.68	0.6	0.67	0.6	0.66	No
CV8	0.74	0.64	0.83	0.72	0.65	Yes
CV9	0.77	0.7	0.66	0.54	0.73	No
CV10	0.7	0.6	0.72	0.66	0.63	No
CV11	0.77	0.66	0.74	0.62	0.69	No
CV 12	0.74	0.64	0.7	0.59	0.69	Yes
CV13	0.79	0.59	0.7	0.59	0.66	Yes
Cv14	0.74	0.65	0.68	0.54	0.65	Yes
Cv15	0.66	0.58	0.83	0.71	0.62	No
CV16	0.74	0.68	0.8	0.69	0.72	Yes
CV17	0.74	0.7	0.77	0.62	0.73	No
CV18	0.7	0.62	0.68	0.57	0.67	No
CV19	0.74	0.57	0.83	0.65	0.67	Yes
CV20	0.79	0.69	0.82	0.75	0.73	Yes
CV21	0.74	0.69	0.8	0.71	0.73	Yes
CV22	0.66	0.69	0.77	0.62	0.72	No
CV23	0.7	0.65	0.72	0.59	0.67	Yes

CV24	0.72	0.6	0.66	0.56	0.66	Yes
CV25	0.72	0.53	0.64	0.54	0.63	Yes
CV26	0.7	0.58	0.64	0.52	0.63	No
CV27	0.66	0.55	0.74	0.62	0.6	Yes
CV28	0.7	0.62	0.74	0.69	0.66	No
CV29	0.7	0.65	0.74	0.66	0.69	No
CV30	0.72	0.63	0.7	0.6	0.69	No
CV31	0.68	0.62	0.81	0.65	0.64	Yes
CV32	0.7	0.66	0.76	0.66	0.69	Yes
CV33	0.77	0.65	0.76	0.68	0.68	No
CV34	0.72	0.71	0.79	0.66	0.69	Yes
CV35	0.72	0.7	0.8	0.65	0.7	Yes
CV36	0.85	0.68	0.64	0.52	0.73	No
CV37	0.66	0.54	0.67	0.54	0.6	No
CV38	0.79	0.58	0.67	0.56	0.64	No
CV39	0.72	0.58	0.66	0.6	0.63	No
CV40	0.81	0.62	0.66	0.61	0.65	No
CV41	0.72	0.65	0.64	0.53	0.62	No
CV42	0.72	0.56	0.75	0.73	0.62	No
CV43	0.85	0.73	0.78	0.7	0.72	No
CV44	0.87	0.73	0.69	0.63	0.78	No

2. The result by traditional LSA approach(similarity by word frequency LSA(word))

Job.posting CVs	J.p1	J.p2	J.p3	J.p4	J.p5	expert
CV1	0.68	0.91	0.66	0.79	0.93	No
CV2	0.32	0.47	0.71	0.55	0.47	Yes
CV3	0.57	0.61	0.84	0.78	0.47	Yes
CV4	0.28	0.41	0.64	0.61	0.25	No
CV5	0.65	0.64	0.85	0.8	0.5	No
CV6	0.53	0.87	0.69	0.83	0.83	Yes
CV7	0.44	0.51	0.69	0.6	0.33	No
CV8	0.29	0.54	0.61	0.62	0.35	No
CV9	0.39	0.49	0.7	0.71	0.36	Yes
CV10	0.86	0.84	0.88	0.79	0.67	No
CV11	0.78	0.8	0.93	0.86	0.69	No
CV 12	0.74	0.76	0.93	0.88	0.65	No
CV13	0.87	0.47	0.69	0.66	0.31	Yes
Cv14	0.63	0.75	0.84	0.87	0.62	No
Cv15	0.69	0.66	0.75	0.74	0.47	No
CV16	0.74	0.35	0.48	0.75	0.48	No
CV17	0.41	0.48	0.73	0.74	0.38	No
CV18	0.48	0.62	0.71	0.48	0.45	Yes
CV19	0.38	0.46	0.71	0.67	0.31	Yes
CV20	0.79	0.73	0.94	0.9	0.63	Yes
CV21	0.14	0.27	0.83	0.77	0.4	Yes

CV22	0.62	0.53	0.42	0.68	0.26	No
CV23	0.5	0.53	0.8	0.75	0.41	Yes
CV24	0.73	0.58	0.67	0.61	0.41	Yes
CV25	0.74	0.85	0.84	0.83	0.54	Yes
CV26	0.81	0.97	0.87	0.88	0.87	Yes
CV27	0.72	0.93	0.83	0.84	0.74	Yes
CV28	0.53	0.53	0.69	0.63	0.41	No
CV29	0.46	0.53	0.7	0.73	0.39	Yes
CV30	0.73	0.68	0.9	0.82	0.53	Yes
CV31	0.46	0.52	0.77	0.81	0.53	No
CV32	0.49	0.53	0.69	0.73	0.39	No
CV33	0.45	0.51	0.72	0.62	0.46	Yes
CV34	0.49	0.62	0.77	0.66	0.51	Yes
CV35	0.68	0.66	0.85	0.77	0.49	Yes
CV36	0.65	0.61	0.9	0.88	0.59	Yes
CV37	0.71	0.51	0.7	0.67	0.35	Yes
CV38	0.77	0.64	0.69	0.82	0.54	Yes
CV39	0.81	0.66	0.93	0.83	0.59	No
CV40	0.7	0.7	0.87	0.82	0.54	No
CV41	0.7	0.45	0.52	0.42	0.4	No
CV42	0.38	0.8	0.43	0.61	0.75	Yes
CV43	0.75	0.61	0.87	0.74	0.51	No
CV44	0.86	0.75	0.94	0.84	0.67	No

3. The result of similarity between CVs and job posting by LSA (concepts only)

Job.posting CVs	J.p1	J.p2	J.p3	J.p4	J.p5	expert
CV1	0.78	0.89	0.83	0.25	0.37	Yes
CV2	0.64	0.34	0.77	0.67	0.56	Yes
CV3	0.83	0.53	0.79	0.86	0.82	No
CV4	0.79	0.63	0.78	0.85	0.87	No
CV5	0.77	0.57	0.74	0.79	0.88	No
CV6	0.88	0.85	0.89	0.66	0.68	No
CV7	0.87	0.82	0.86	0.82	0.83	No
CV8	0.9	0.85	0.89	0.8	0.79	No
CV9	0.39	0.42	0.39	0.34	0.7	No
CV10	0.89	0.82	0.9	0.69	0.73	No
CV11	0.79	0.76	0.78	0.76	0.77	No
CV 12	0.64	0.4	0.63	0.56	0.83	No
CV13	0.89	0.79	0.88	0.85	0.84	Yes
Cv14	0.87	0.78	0.86	0.86	0.88	Yes
Cv15	0.85	0.72	0.83	0.82	0.84	No
CV16	0.8	0.51	0.77	0.85	0.79	No
CV17	0.77	0.43	0.73	0.85	0.85	No
CV18	0.52	0.19	0.47	0.84	0.63	No
CV19	0.80	0.61	0.77	0.86	0.82	No
CV20	0.7	0.38	0.66	0.83	0.74	No
CV21	0.64	0.26	0.59	0.80	0.86	No
CV22	0.82	0.5	0.78	0.87	0.78	No
CV23	0.71	0.38	0.66	0.85	0.71	No
CV24	0.78	0.45	0.74	0.87	0.75	No

CV25	0.78	0.83	0.88	0.77	0.87	No
CV26	0.79	0.90	0.82	0.44	0.52	Yes
CV27	0.87	0.86	0.89	0.62	0.68	No
CV28	0.88	0.64	0.86	0.83	0.86	No
CV29	0.78	0.75	0.86	0.88	0.86	Yes
CV30	0.88	0.57	0.85	0.83	0.86	No
CV31	0.8	0.48	0.76	0.87	0.76	No
CV32	0.64	0.3	0.6	0.80	0.76	No
CV33	0.82	0.49	0.79	0.86	0.88	No
CV34	0.61	0.25	0.57	0.83	0.88	No
CV35	0.88	0.75	0.86	0.89	0.85	No
CV36	0.86	0.6	0.83	0.84	0.8	No
CV37	0.79	0.78	0.87	0.86	0.86	No
CV38	0.79	0.83	0.79	0.89	0.83	Yes
CV39	0.79	0.8	0.78	0.87	0.81	Yes
CV40	0.89	0.81	0.88	0.84	0.84	No
CV41	0.88	0.81	0.87	0.89	0.85	Yes
CV42	0.7	0.85	0.75	0.65	0.51	Yes
CV43	0.87	0.72	0.85	0.9	0.87	Yes
CV44	0.89	0.79	0.87	0.85	0.8	No

4. The result for our approach

Job.posting CVs	J.p1	J.p2	J.p3	J.p4	J.p5	Expert
CV1	0.42	0.96	0.62	0.4	0.83	Yes
CV2	0.54	0.65	0.75	0.53	0.55	Yes
CV3	0.62	0.63	0.94	0.55	0.72	Yes
CV4	0.86	0.6	0.73	0.86	0.87	No
CV5	0.74	0.6	0.73	0.75	0.57	Yes
CV6	0.62	0.94	0.74	0.61	0.88	Yes
CV7	0.64	0.56	0.58	0.51	0.74	Yes
CV8	0.68	0.82	0.78	0.65	0.85	No
CV9	0.82	0.35	0.52	0.8	0.36	No
CV10	0.7	0.85	0.75	0.69	0.84	Yes
CV11	0.61	0.53	0.6	0.48	0.78	Yes
CV 12	0.84	0.61	0.74	0.8	0.65	Yes
CV13	0.82	0.71	0.81	0.73	0.73	Yes
Cv14	0.76	0.71	0.84	0.73	0.74	No
Cv15	0.78	0.6	0.74	0.77	0.62	No
CV16	0.62	0.52	0.77	0.6	0.55	Yes
CV17	0.86	0.27	0.84	0.14	0.42	Yes
CV18	0.45	0.39	0.9	0.35	0.52	Yes
CV19	0.53	0.55	0.91	0.46	0.64	Yes
CV20	0.44	0.44	0.92	0.35	0.54	Yes
CV21	0.52	0.44	0.93	0.42	0.54	Yes
CV22	0.75	0.47	0.73	0.48	0.5	Yes
CV23	0.46	0.38	0.84	0.39	0.48	Yes
CV24	0.77	0.51	0.51	0.48	0.49	Yes

CV25	0.79	0.87	0.7	0.81	0.7	Yes
CV26	0.7	0.97	0.73	0.68	0.93	Yes
CV27	0.75	0.89	0.81	0.72	0.87	Yes
CV28	0.92	0.42	0.5	0.95	0.45	Yes
CV29	0.80	0.69	0.73	0.82	0.69	Yes
CV30	0.62	0.6	0.88	0.58	0.65	Yes
CV31	0.51	0.48	0.88	0.43	0.59	Yes
CV32	0.58	0.38	0.88	0.49	0.51	Yes
CV33	0.47	0.4	0.92	0.36	0.53	Yes
CV34	0.44	0.34	0.89	0.33	0.45	Yes
CV35	0.67	0.69	0.87	0.63	0.74	Yes
CV36	0.67	0.6	0.9	0.6	0.67	Yes
CV37	0.76	0.73	0.75	0.73	0.72	Yes
CV38	0.95	0.54	0.51	0.97	0.56	Yes
CV39	0.88	0.68	0.76	0.85	0.74	No
CV40	0.71	0.75	0.73	0.68	0.75	Yes
CV41	0.91	0.32	0.28	0.9	0.34	No
CV42	0.74	0.92	0.57	0.7	0.87	Yes
CV43	0.86	0.65	0.7	0.88	0.66	Yes
CV44	0.54	0.55	0.62	0.43	0.79	Yes