Fall 2011

Enhancing Productivity of Recruitment Process Using Data Mining & Text Mining Tools

Charul Saxena
San Jose State University

Enhancing Productivity of Recruitment Process Using Data mining & Text Mining Tools

A Project Report

Presented to

The Faculty of the Department of Computer Science

San Jose State University

In Partial Fulfillment of the Requirements for the Degree

Master of Computer Science

By

Charul Saxena

Fall 2011

The Undersigned Project Committee Approves the Project Titled

Enhancing Productivity of Recruitment Process Using

Data mining & Text Mining Tools

By

Charul Saxena

APPROVED FOR THE DEPARTMENT OF COMPUTER SCIENCE

_____

Dr. T. Moh, Department of Computer Science Date

_____

Dr. Chris Pollett, Department of Computer Science Date

_____

Dr. Mark Stamp, Department of Computer Science Date

APPROVED FOR THE UNIVERSITY

_____

Associate Dean Office of Graduate Studies and Research Date

ABSTRACT

ENHANCING PRODUCTIVITY OF RECRUITMENT PROCESS USING DATA MINING & TEXT MINING TOOLS

By

Charul Saxena

Digital communication has significantly reduced the time it takes to send a résumé, but the recruiter's work has become more complicated because with this technological advancement they get more résumés for each job opening. It becomes almost impossible to physically scan each résumé that meets their organization's job requirement. The filtering and search techniques provide hundreds of résumés that can fulfill the desired criteria. Most approaches focus on either parsing the résumé to get information or propose some filtering methods. Moreover, résumés vary in format and style, making it difficult to maintain a structural repository which would contain all the necessary information.

The goal of this project is to examine and propose an approach which would consider the skill sets from the potential résumés, along with expertise domains like related work experience and education, to score the selected "relevant résumé." This approach aims at highlighting the most important and relevant résumés, thus saving an enormous amount of time and effort that is required for manual scanning by the recruiters.

The study presented here is based on the real world data-set of résumés. It indicates that the proposed idea has the potential to improve the process used to select résumés and highlight the key features of each candidate, and draw attention to the key skills required for a specific job.

ACKNOWLEDGEMENTS

I would like to express my deep and sincere gratitude to my project advisor, Dr. T. Moh for his guidance, encouragement, and support throughout the project. I extremely appreciate Dr. Mark Stamp and Dr. Chris Pollett for their input and support. I would like to thank my committee members who helped me in improving my thesis.

I would also like to extend sincere thanks to my friends Satyam, Dhiti, Gaurav, Pooja and Ashwani who helped me to collect the real time data for my research project and contributed valuable suggestions about my thesis. Last but not the least, I would like to thank my parents and sister who supported and encouraged me at every step.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

All major industries today are driven by technology. According to current statistics, information available on the internet is about 60% of what we need [1]. This figure is expected to rise exponentially in the near future. Companies are publishing more and more information on the internet about every aspect of their business and their growth [2].

Recruiters receive large numbers of applications through e-mails, online job portals, or through services provided by partner staffing companies [3]. Online job portals like monster.com, indeed.com, dice.com, and careerbuilder.com and staffing firms like Manpower Inc., Adecco, and Kelly Services draw in most of the applications [4].

Résumés obtained from such diverse sources are thus difficult to process and store in a unified database format. It becomes very tedious to select the most appropriate ones. Since résumés are structured documents containing information based on the author's thinking and writing skills, they can be created in a multitude of formats (e.g., plain text or structured table), languages and file types (e.g., txt, pdf, and doc.). This makes the information extraction (IE) process highly complex. High precision and recall becomes complicated for this domain.

Dynamic filtering techniques are used by the industry to extract relevant résumés. These filtering techniques match hundreds of résumés from the database to a single job posting. Résumés extracted by these filters are generally similar to each other as they satisfy the same search criteria, based mainly on keyword matching. The résumé filtering becomes more challenging

when the job requirement demands a specialized skill set. Thus, it becomes cumbersome to further analyze the short-listed résumés in order to select the most relevant résumés.

1.1 PROBLEM STATEMENT

The online job portals, staffing agencies, and recruiters deploy filtering techniques or search services to obtain a few hundred potential candidates' profile from terabytes of data present in the résumé database. The hiring managers/human resource (HR) business partners obtain the segregated set of résumés which are similar to each other. Next, a manual analysis of each résumé is required to mine for the best candidates. This is referred to as "Problem of Résumé Selection" [5].

The search keywords entered by recruiters into a major job board, a total of 3,004 times in the December of 2010, were "occupational therapist," "certified occupational therapist," "cota," "physical therapist," "speech language pathologist," and "speech-language pathologist" [6]. A keyword search like this results in a large number of résumés, from the repository, which contain these particular keywords and disregard the considerably important information present in the résumé.

There are many approaches which focus on identifying and extracting information from résumés, some of which focus on storing the information obtained from the résumés in a structured format. Very little research has been done on providing the best match for a particular

requirement. Recruiters have to scan all the similar looking résumés manually, after applying the filters.

1.2 CONTRIBUTION OF THESIS

The objective of this research is to propose an algorithm that provides a precise list of prospective candidates with relevant experience and then present the highlights of each selected résumé. The result of this algorithm would provide an overview of skills for each profile, for comparing.

In this paper, the keyword based approach takes into account the specific specialization in a certain technical field, along with the relevant skills, experience and education. Additionally, this method also involves extracting the special information from the résumé and organizing it efficiently, in order to enhance and optimize the résumé selection process. Adoption of this method should significantly lessen the manual processing time required to scan through the relevant résumés with similar skill sets, thus reducing the workload for the recruiters and HR managers to choose the best résumé from the set. This problem can be termed as “Extraction of Relevant Résumés.”

The approach proposed here aims to order the similar résumés obtained after the application of various filters. Some of the major contributions of the algorithm are:

1. Providing a ranking-based approach for the similar résumés that are obtained after filtering.

2. Proposing a framework, to highlight the unique skills of a résumé, with 3-tier architecture.
3. Comparison of features of the résumés with other prospective candidate profiles.

1.3 ORGANIZATION OF REPORT

Section 2 provides the background of the résumés and different aspects related to the recruitment industry. The section briefly explains the problems being faced by the HR and hiring managers in résumé selection. Section 3 provides an overview of the concepts used in this project. In Section 4, the related work in this domain is described that lays and strengthens the foundation of my approach. The proposed approach with examples is discussed in Section 5. It also highlights the design and implementation details of the résumé selection process which also considers the relevant information from the résumé. Section 6 provides the experimental results and the comparison of the existing approaches, followed by the conclusion in Section 7. Finally, Section 8 provides some recommendations for possible future work.

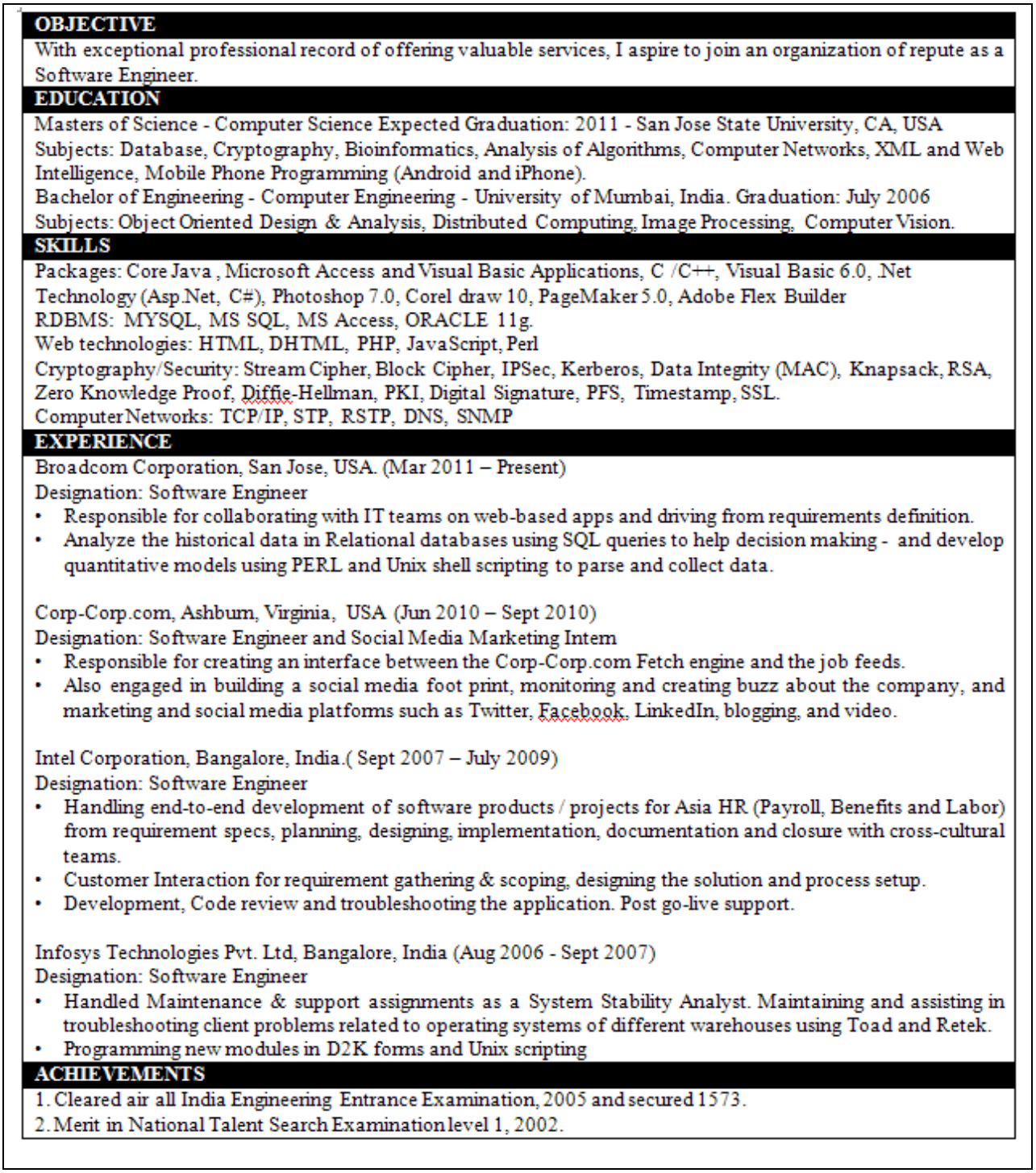
2. BACKGROUND

A résumé is a brief document about an individual trying to market him/her to the industry. They usually are structured and hierarchical documents but contain unstructured data too [7]. In a résumé, the format is not predetermined and it is based on the author's thinking, which makes the information extraction, comparison, and selection a daunting task. Each résumé is unique in its own way as it contains words and sentences as features [8]. It can also be viewed as a multi-section document, with description of each section, which highlights the different aspects of an individual's professional career [9].

The layered structure in a résumé usually consists of two parts: *Section Header* and *Section Information*. Both the parts are related to each other and appear in the same textual block. Usually résumés have multiple sections of two-layered architecture. Education, work experience, relevant skills and personal information are all examples of a résumé section. Each is a part of the structured layer and their description forms the unstructured layer in the above mentioned two-layered architecture. A sample of the layered résumé structure is shown below in Figure 1, page 6.

Traditionally, résumés on high quality parchment paper were used to highlight a person's accomplishments, knowledge, skills, abilities, and experiences [10]. The advent of technology is growing in every industry [11], and the recruitment industry is not an exception to it. It is reaping the benefits of this technological era. The recruitment industry is going paperless and preferring the electronic résumés compared to the paper ones. Paper résumés require a considerable amount

of manual work which is not the case with the electronic format. The electronic résumés are far more efficient, convenient to store, quicker to edit, and can be easily accessed later.



OBJECTIVE

With exceptional professional record of offering valuable services, I aspire to join an organization of repute as a Software Engineer.

EDUCATION

Masters of Science - Computer Science Expected Graduation: 2011 - San Jose State University, CA, USA
Subjects: Database, Cryptography, Bioinformatics, Analysis of Algorithms, Computer Networks, XML and Web Intelligence, Mobile Phone Programming (Android and iPhone).
Bachelor of Engineering - Computer Engineering - University of Mumbai, India. Graduation: July 2006
Subjects: Object Oriented Design & Analysis, Distributed Computing, Image Processing, Computer Vision.

SKILLS

Packages: Core Java , Microsoft Access and Visual Basic Applications, C /C++, Visual Basic 6.0, .Net Technology (Asp.Net, C#), Photoshop 7.0, Corel draw 10, PageMaker 5.0, Adobe Flex Builder
RDBMS: MYSQL, MS SQL, MS Access, ORACLE 11g.
Web technologies: HTML, DHTML, PHP, JavaScript, Perl
Cryptography/Security: Stream Cipher, Block Cipher, IPSec, Kerberos, Data Integrity (MAC), Knapsack, RSA, Zero Knowledge Proof, Diffie-Hellman, PKI, Digital Signature, PFS, Timestamp, SSL.
Computer Networks: TCP/IP, STP, RSTP, DNS, SNMP

EXPERIENCE

Broadcom Corporation, San Jose, USA. (Mar 2011 – Present)
Designation: Software Engineer

- Responsible for collaborating with IT teams on web-based apps and driving from requirements definition.
- Analyze the historical data in Relational databases using SQL queries to help decision making - and develop quantitative models using PERL and Unix shell scripting to parse and collect data.

Corp-Corp.com, Ashburn, Virginia, USA (Jun 2010 – Sept 2010)
Designation: Software Engineer and Social Media Marketing Intern

- Responsible for creating an interface between the Corp-Corp.com Fetch engine and the job feeds.
- Also engaged in building a social media foot print, monitoring and creating buzz about the company, and marketing and social media platforms such as Twitter, Facebook, LinkedIn, blogging, and video.

Intel Corporation, Bangalore, India.(Sept 2007 – July 2009)
Designation: Software Engineer

- Handling end-to-end development of software products / projects for Asia HR (Payroll, Benefits and Labor) from requirement specs, planning, designing, implementation, documentation and closure with cross-cultural teams.
- Customer Interaction for requirement gathering & scoping, designing the solution and process setup.
- Development, Code review and troubleshooting the application. Post go-live support.

Infosys Technologies Pvt. Ltd, Bangalore, India (Aug 2006 - Sept 2007)
Designation: Software Engineer

- Handled Maintenance & support assignments as a System Stability Analyst. Maintaining and assisting in troubleshooting client problems related to operating systems of different warehouses using Toad and Retek.
- Programming new modules in D2K forms and Unix scripting

ACHIEVEMENTS

1. Cleared air all India Engineering Entrance Examination, 2005 and secured 1573.
2. Merit in National Talent Search Examination level 1, 2002.

Figure 1: Sample résumé having hierarchical structure with sections and their respective description

The staffing departments of many organizations have concluded that the internet is an inexpensive and more visible platform for posting jobs. It costs $200 to $400 to post the job on the internet compared to the expensive and less evident print media which can cost anywhere between $3000 to $7000 for a one-time advertisement [4]. A recent survey found that about 78% of the employers were content with investing in the online recruiting domain [4].

The résumés in the online format are scanned and processed using information extraction methods, which are then stored in the employer's résumé repository [12]. When a recruiter requires some résumés, he/she performs a keyword search on this résumé database. This search uses dynamic filtering techniques. The set of filters help in tossing out irrelevant résumés [13]. Google uses "résumé filters" to sift through the piles of résumés it gets every month for every job posted. Online job applicants provide some personal information that helps the recruiter to delve into factors such as attitude, behavior, and personality traits. Computations are performed based on these factors to arrive at a score that determines how well a person would fit into the Google "culture" [14]. The filtering programs vary significantly; the simpler filters may provide very basic assistance to a company's human resource department whereas the more complex ones may employ sophisticated techniques such as those used by Google. Irrespective of which mechanism is applied, the trend of using filters is growing tremendously.

Experts and recruiters state that using résumé filters purges around 60 to 75% of the unwanted documents. However, applying strict filters to a set might adversely affect the search [14]. Enforcing exceptional search criteria may result in certain prospective candidates' résumés being ignored. Even though strict filters might neglect certain potential résumés, the speed and

accuracy with which the results are computed are so impressive that the recruiters and hiring managers have started to favor electronic methods for determining suitable candidates [9].

Electronic media usage has definitely benefitted the recruiters and employers because specific criteria and objective measures like skills, education, experience, industry, age, qualification, availability, discipline, and location can be specified and extracted from the résumés [9]. The keywords signify the unique features that an employer is seeking in a potential candidate. To narrow the search further, employers can re-filter the shortlisted résumés by setting more precise criteria such as the school name - “Ivy League school education,” position title - “Sales Manager position,” current employer - “IBM,” technologies applicable - “Microsoft Message Queuing (MSMQ) experience,” “networking technology,” and “HR work background” which seem to specifically relate to the requirements. Re-filtering would fine-tune the search and provide the recruiter with a set of limited but prospective candidates. The technique of filtering and then specifying the critical criteria vigilantly provides a cluster of the most appropriate candidates by eradicating the profiles which do not possess the desired qualifications.

However, the task of information and special features extraction from a résumé is not as straight forward as it seems. Each résumé consists of several sub-topics, and each of those may contain text in different formats. For example, the experience sub-section may be composed of long sentences in free-form text, the education sub-section usually describes the schools attended and the degrees obtained as a bulleted list whereas the skills sub-section contains skill categories (e.g., programming languages) and specific skills (e.g., C++, Java). Hence, the information extraction from the résumé dataset is a complex task. Each sub-section needs a different

approach for the varied data content. The main challenge is to divide the sub-sections into a set of features with a résumé uniqueness framework. For example, with the keyword search, we would obtain a cluster of similar candidate résumés for a single domain like computer science. One may note that the résumés from the same domain will have lot of common features like skills, tools, and technologies. Along with these attributes, each résumé may possess a multitude of special skills thus differentiating one document (candidate) from the others [5]. Hence, the procedure of extracting the special skills from each prospective candidate's résumé would help employers, interviewers, and recruiters to efficiently shortlist the most applicable résumés.

Special information may exist in some résumés which might provide a better insight in making a better selection. These résumés may have special achievements in diverse sections like education, experience, and skills. Thus, identifying and extracting such distinguishing features of a résumé and organizing them effectively to provide a concise summary helps speed up the résumé selection process [9].

In this paper, an approach to rank and prioritize these résumés has been provided. Once these résumés are ranked, an efficient method of highlighting the “unique” features of a résumé in a set provides an easy way of sifting through the résumé stack.

3. OVERVIEW

Internet usage has increased exponentially in the arena of job search. This has provided everyone in the market with a better opportunity to access not only the jobs but also the résumés. It is not uncommon for hundreds, if not thousands of applicants to apply for a single job [14]. But, with the spread of technology, each prospective candidate also has to cross numerous hurdles of technology. Each résumé that is posted gets accessed and scanned for the details. Technological algorithms determine whether a particular job application gets to go to the next step or needs to be tossed out. Then eventually the résumés obtained are manually reviewed by recruiters and hiring managers to shortlist the most apt candidates.

3.1 INFORMATION EXTRACTION

Information Extraction (IE) is a kind of Information Retrieval method used to automatically extract structured information from a large collection of unstructured documents. The output of the application of IE varies; however, it can be used to populate a database of interest.

An IE system extracts the relevant information from a set of data with known information types. For example, if a user wants to compare the resolutions of two digital cameras, Sony A390 and Sony A560, the keywords required to get this information would be “Sony A390” and “Sony A560”. An IE system would thus be required to have a set of documents containing information about the camera names and their resolutions. This information is extracted and provided in such a manner that the features can be listed and compared [5].

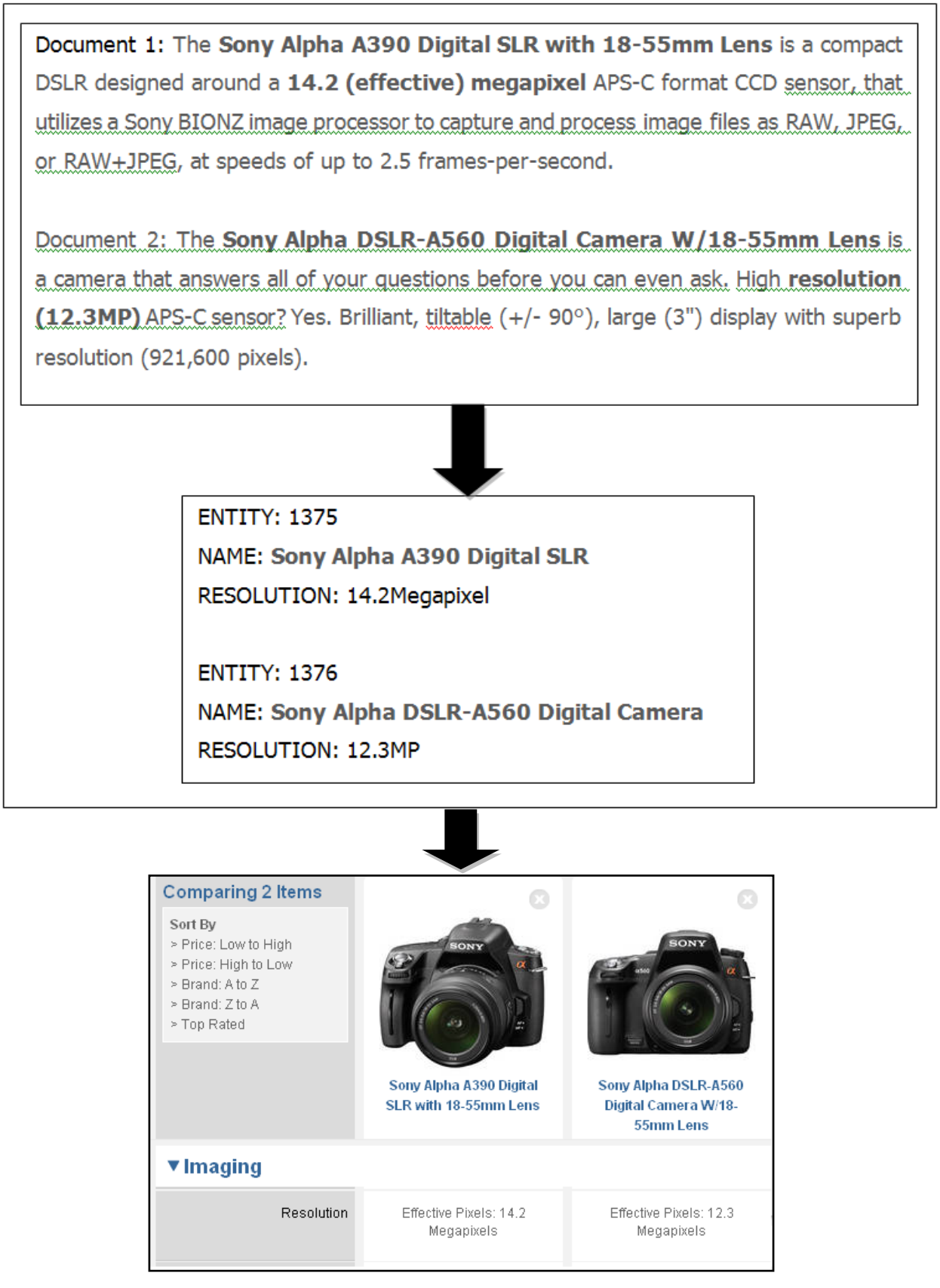


Figure 2: Information Extraction Example

For the purpose of this project, we use “VisualText” analyzer for information extraction. It is an open source tool to develop fast information extraction, natural language processing, and text analysis systems. “The project emphasized development of the automated rule generation (RUG) capability (a prototype capability in VisualText).” This résumé extraction prototype extracts personal contact information, experience, and education information from résumés with 80% accuracy (90% precision and 75% recall/completeness) [15]. This prototype can be modified to obtain the different sections from a hierarchical résumé. The output provides separate files for each section like education, experience, personal information, and skills along with the extracted information in an xml format. This xml file and the separate text files generated are used to populate the data for the experiment, in order to analyze the proposed algorithm.

3.2 NORMALIZATION

One of the biggest advantages of IE is that it can be used to extract information from a huge volume of unstructured data. However, as we are dealing with free-form of text, every document is distinct. In the case of résumés, each résumé is a human-generated text with a different language that has the freedom of choosing the words, format, structure, and content. To compare such a myriad array of documents, we need to provide a platform which would translate all the given documents on the same scale. “The process of mapping the extracted strings to a predefined format is called *Normalization* [1].” The information extracted units are required to be normalized in order to map them to an existing database or to compare their values. In a nutshell, normalization or rescaling is performed to translate values in different ranges to the same scale.

We perform the normalization on the strings extracted from résumés by comparing canonical names from the database. For example, there are different ways to mention a specific technology in a résumé – “shell,” “Shell scripting,” or “shell script.” All these names can be mapped to “Shell.” In addition to this, the scoring methodology proposed normalizes the values to scale the scores of the different sections on the same platform.

3.3 CLUSTERING

Clustering can be defined as the process of creating clusters. Each cluster is a collection of objects which are similar in some manner. It usually deals with finding a similarity in an unstructured collection of unlabeled data.

The K-Means Nearest Neighbor algorithm is a machine-learning instance-based technique. This method does not construct models but stores the training instances. For each new instance, the algorithm compares the distance feature-vectors to the training set. The nearest neighbors are selected based on the distance of the features of the new instance i.e., the similarity between the new instance and the training set vectors. “K” in the algorithm defines the number of nearest neighbors. The classification of the new object is based on the distance between K-clusters and the object. The object is assigned to the cluster with the minimum distance.

This algorithm is iterative in nature and repeats for each object. It converges until the objects are stable (i.e., no object changes in the group) [16]. K-Means clustering is simple, and the basic steps it follows are:

1. Number of clusters, K, is determined.
2. Assume a centroid or center of the K clusters. Any object can be randomly chosen and initialized as an initial centroid, or the first K objects can also serve as the initial centroids.
3. The distance of each object from each of the centroids is calculated.
4. Group the objects based on minimum distance (find the closest centroid for each object)

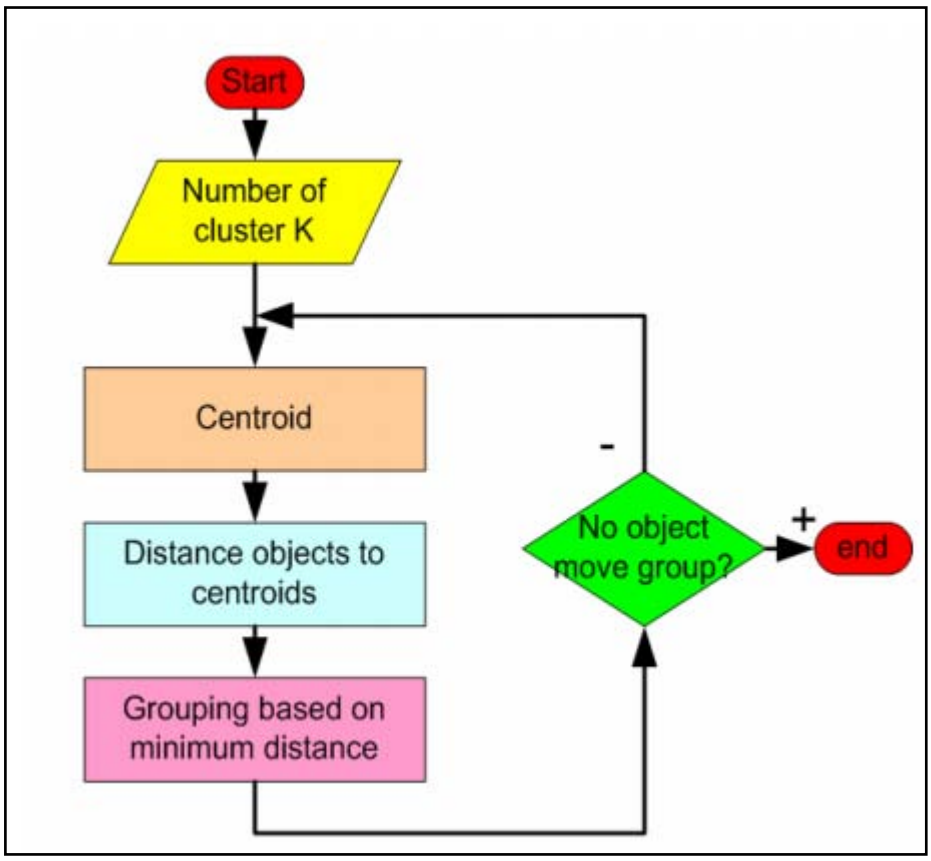


Figure 3: K-Means clustering algorithm

We use **strict partitioning clustering** to group our résumés where each résumé belongs to exactly one cluster. Furthermore, based on the experience of candidates in the data-set, K-Means clustering is used to divide the selected profiles into two clusters as we need two partitions from the selected profiles. These are made of a single set containing the résumés close to minimum experience and the other close to maximum experience in the set. To display the unique features of a résumé and its similarity to other résumés in the set, we use simple clustering based on the skills present in a résumé data-set.

3.4 SKILL CATEGORY, SPECIFIC SKILL, AND SPECIAL (UNIQUE) SKILLS

Résumés can be distinguished based on different criteria like experience, education and location. However, the main section of the résumé, that would help distinguish each document from the rest, is the skills section. Each job posting has some special requirements defined in terms of the skills that an employer is looking for, in a candidate. This specific requirement can be handled by extracting the special information available in the résumé, which is in the form of “skills”. The skills which can be considered can be compared to features of a product, such as the digital camera in Figure 2, page 11. The cameras under consideration may share some common features and some distinguishable features which would help the users decide while choosing the product [15]. We can extend this notion to the skill set of résumés under observation.

We have assumed that skills follow a hierarchical pyramid structure. The *Skill Category* defines the broad domains like Database, Web Programming, and Mobile Platforms forming the base of the pyramid. Each domain contains specific skills. For example, MySQL, MS Access, and

Oracle are specific database skills. This forms the second tier of the pyramid. This level is useful for a basic keyword search that determines similar sets of résumés.

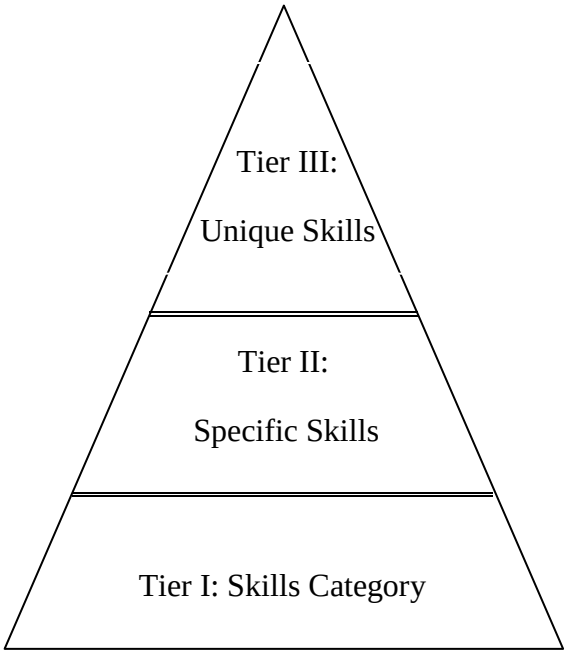


Figure 4: Skill Pyramid

The *Unique skills* that form the third and top most tier of the pyramid are important to determine the uniqueness of a résumé (candidate). This level can be visualized as a subset containing the most appropriate résumés from the set which highlighted the specific skills.

4. RELATED WORK

Extraction of the information from résumés has been an important area of focus for a lot of researchers. The work on résumés usually involves information extraction parsers, classifiers, and natural language processors, and structures to store the data.

There are many commercial products on résumé data storing, information extraction and retrieval, but there has been very limited published research work in this area. Some of the commercial products include: Daxtra CVX [17], Sovren Résumé/CV Parser [18], ALEX Résumé parsing [19], Akken Staffing [20], and RésuméGrabber Suite [21]. The product specification, algorithms and methods used in them for résumé information extraction are not available completely.

4.1 RÉSUMÉ INFORMATION EXTRACTION

One of the published studies tried to learn the information extraction rules for résumés written in English using an adaptive transformation based toolkit called "Learning Pinocchio $(LP)_2$." The system performs the IE by annotating texts using XML tags to identify elements such as name, street, city, province, email, etc. [22].

Another approach applied the concept of information retrieval to extract information from online Chinese résumés where regular expression and text automatic classification were used to extract

basic information from a résumé while fuzzy logic algorithm was used to extract the complex information [23].

Many résumé information extraction systems employ a hybrid approach by using a combination of different methods. An approach examined cascaded two-pass IE framework that used Hidden Markov Model (HMM) and Support Vector Machines (SVM) which are statistical and learning-based methods respectively. In the first pass, HMM segments rèsumès into blocks that represent different types of information. In the second pass, HMMs and SVMs extract general information from the annotated blocks, with different classifiers trained to extract different types of information [3].

4.2 RÉSUMÉ STORAGE

Résumé IE includes structuring, grouping, and preparing unstructured data to populate a database. There are many approaches proposed to support automatic résumé management and routing by résumé information extraction [24]. A four phase approach has also been proposed that processed the résumés to produce the extracted information in JSON or XML format [25].

4.3 RÉSUMÉ FILTERS AND CANDIDATE PROFILERS

The filtering software like Applicant Tracking Systems (ATS) scans the résumés and identifies the key phrases that could deliver a set of candidate résumés which might be most promising. The application of filters on the résumés, help the recruiters to wade through the large volume of

résumé from data box. ATS uses sophisticated screening and sorting functions and parsers to help sort and categorize résumés. Usually, the filtered résumés are presented in the order of the most recently submitted ones on the top. However, these filtering techniques do not consider the special skills, related and relevant experience. They use the keyword and logical searches to provide sets of résumés matching job criteria [6].

RésuméGrabber Suite is a selection tool that captures information from various sources such as job boards, emails attachments, online résumés, and stores them into a database. The suite uses this stored information to provide assistance to employers [21]. Talentdrive.com, software based talent/résumé search and evaluation tool, selects and conducts an analysis from various online résumé databases to provide suggestions for ideal candidates [26]. Yet another tool, Résumé Dragon is also known for similar functions with extras such as background checks and drug screening [27].

Résumé Manager Pro from sarmsoft.com is a résumé filter [10]. Candidate profile search engines such as monster.com and dice.com use the extracted information from résumés like location, job-title, career level, job category, and more to provide list of qualified résumés. The list provided consists of similar résumés as they fulfill certain search criteria [21].

4.4 OTHER RELATED WORK

An e-commerce algorithm investigates the setback faced by customers during ‘selecting a product from a group of similar products.’ The authors use the common features of the similar

products to highlight the uniqueness and help in decision making. This approach can be used as a reference for multi-layered résumés structures [8].

Follow-up work by Yi et al. used the structured relevance model for matching résumés to recruiter queries. The work compares the standard approach to the Structured Relevance Models using the résumé and job requirement dataset as samples. The paper concludes that only 20% accuracy is obtained by pairing unstructured data using relevance models [12].

Some research has also been done to identify the potential benefits and challenging issues faced in using the learning system for recruitment and HR modules [28].

In this paper, we intend to suggest a refined approach to provide relevant résumés obtained after a criteria search by unique features and specialization area.

4.5 DIFFERENCE OVER EXISTING APPROACH

The approach proposed combines the selection and the easy display of the selected set of résumés. Compared to [12], we use keyword matching and normalization to map a job requirement with prospective candidates.

Sumit, Abhishek, and P. Krishna use set of similar résumés for mapping them to a job requirement and then highlight the uniqueness for easy selection. This approach considers only the skills section of the résumés to determine the "*Degree of Specialness*." Authors have made an

attempt to categorize the skills in a résumé by clustering them on the basis of skill types or skill values which helps to select and provide relevant rèsumès [8]. On the contrary, approach proposed in this report parses full rèsumè and searches for the skill values and skill types from a pre-defined database. This technique will not overlook any skills mentioned in other sections of the rèsumè. In addition, the algorithm in proposed approach ranks the rèsumè to provide a "*Rèsumè Relevance Index.*" The rèsumès are further categorized to determine their distinctiveness when compared to others.

This approach provides better results and reduction factor as compared to [8] as the recruiters get relevant rèsumès matching significant requirement criteria with highlight to résumé's special features. An example would be to find a rèsumè for a job requirement of a Siebel Developer with 5 years of experience and a minimum education criteria as Master's degree as education criteria. All the rèsumès with the "Siebel" keyword as the skill section will be selected using the approach in [8]. However, the approach proposed selects the rèsumè with "Siebel Developer" as their skill set during education or work experience based on relevant weightage of experience and education specified by the recruiters.

5. APPROACH

The motivation of the proposed approach is based on the observation that for each job requirement, the recruiter needs to be presented with best possible rèsumè matches. The selected rèsumès should highlight critical information such that it takes only a glance to decide the next level prospective candidates.

In this project, we recommend an approach to overcome the issues of mistakenly discarding potential rèsumès. For example, consider a set of 100 rèsumès. A job requirement is to select résumé of candidate with 5 years of professional experience in Java and has a Master's degree. The existing approaches will filter out approximately 50% of the rèsumès when we apply the first filter as Master's degree. Now, only 50 rèsumès would be obtained after filtering and we apply another filter for 5 years of experience. Now our search comes down to 25 rèsumès. Application of another filter brings our search down to 10 rèsumès. These résumé can be considered as prospective or selected résumés matching all the job criteria.

However, this filtering technique selects rèsumès in the first filter that does not have a Master's degree but not the ones which might have more experience. Further, from filtered rèsumès, we select the ones with less than 5 years of experience. This sequence of filters applied to the set affects the rèsumès selected. If filter for candidate with Master's degree is applied first, we might leave some rèsumès with 5 years of relevant experience and vice versa. So, it becomes very important for the approach to actually make a decision about order of filter being applied.

The approach proposed to apply a filter to the entire set of rèsumès, and not filtering the rèsumès at each level. This gives an appropriate set of rèsumès which is in conjunction with the requirement of the search.

The data from the rèsumès is extracted using the VisualText analyzer. It uses a mixture of language processing methods and provides an xml format of the information from the rèsumès. For the purpose of the project, the analyzer was customized to provide text files for each section of a rèsumè. Each sub-section in the following chapter defines the step by step process to analyze rèsumè data. The flowchart in Figure 5 provides a brief overview of the proposed approach and an insight to the flow of search.

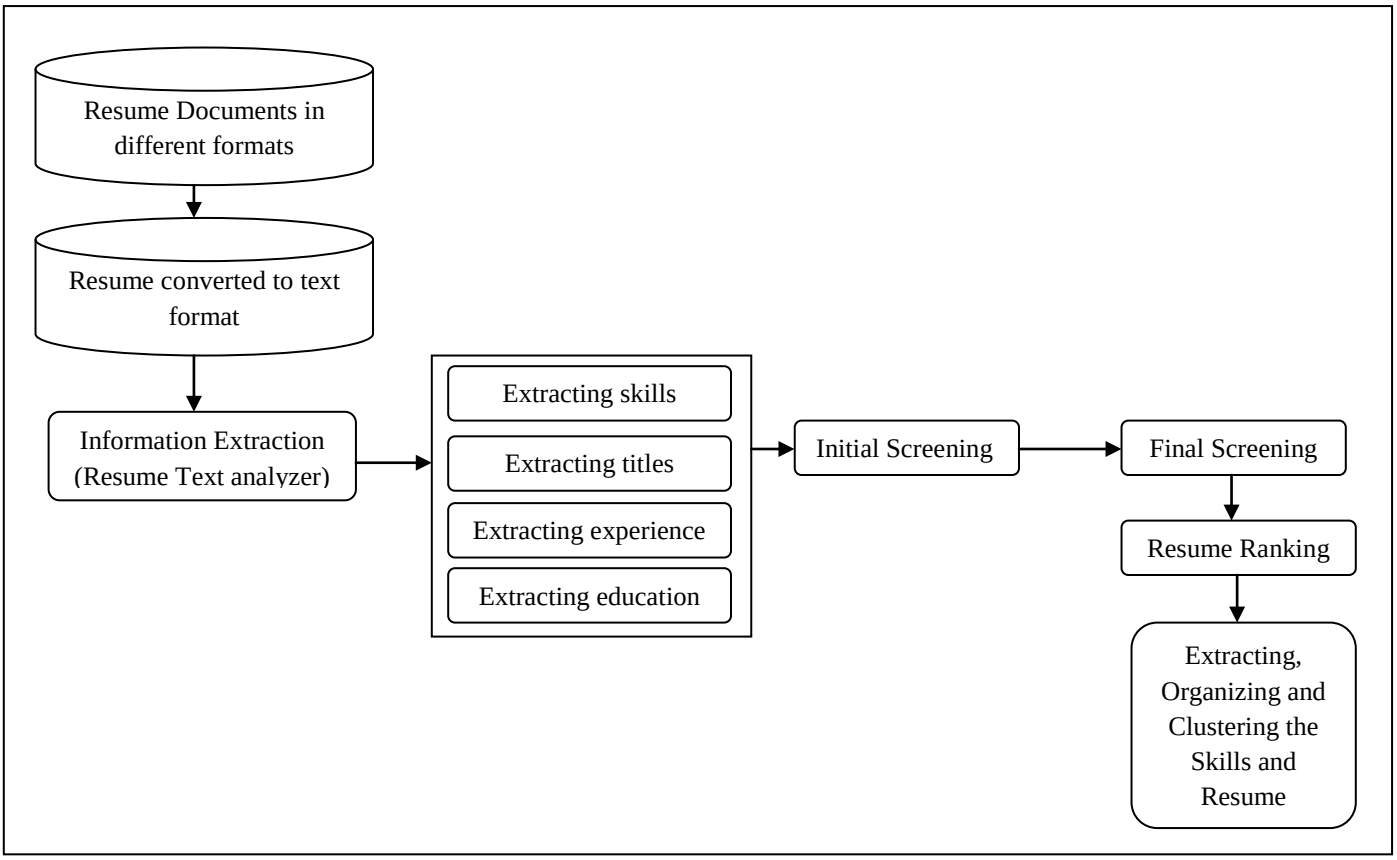


Figure 5: Flowchart of the overall framework

Step 1: Information Extraction from various sub-sections of a Résumé

The desired information from the résumés is extracted in two forms: xml data and sectional text files. Structured data like educational degree, degree duration, professional experience, job title, years of experience, etc., is extracted from the xml file generated. Also, information about the skills categories, specific skills and special skills is extracted using the sectional text files generated by the analyzer.

Identifying Skill Category, Specific Skill and Special Skills

Let R represent a set of 'n' similar rèsumès, where rèsumè $r_i \in R$. The skill, education and experience section of a rèsumè r_i are partly composed of a skill category, a specific skill and special skill features. Let a function of r_i, $f1(r_i)$ be a set of specific skills and $f2(r_i)$ be a set of special skill features for rèsumè r_i. Let F1 represent a set of all specific skills and F2 be set of all special skill features.

The specific skills and unique skills are extracted from each rèsumè while the skill category can be determined through a mapping function M. The mapping function M maps the skill category, specific skill, and special skills. For example, refer to the Figure 6, page 25, the rèsumè sample in Figure 1, page 6 and hierarchical prototype of three levels of rèsumè skill set in Figure 5, page 23. From the skills, education, and experience section, determine the set of specific skills S1 mentioned in the rèsumè – Java, Oracle, Siebel, etc., and set of special skills S2 like "Siebel developer."

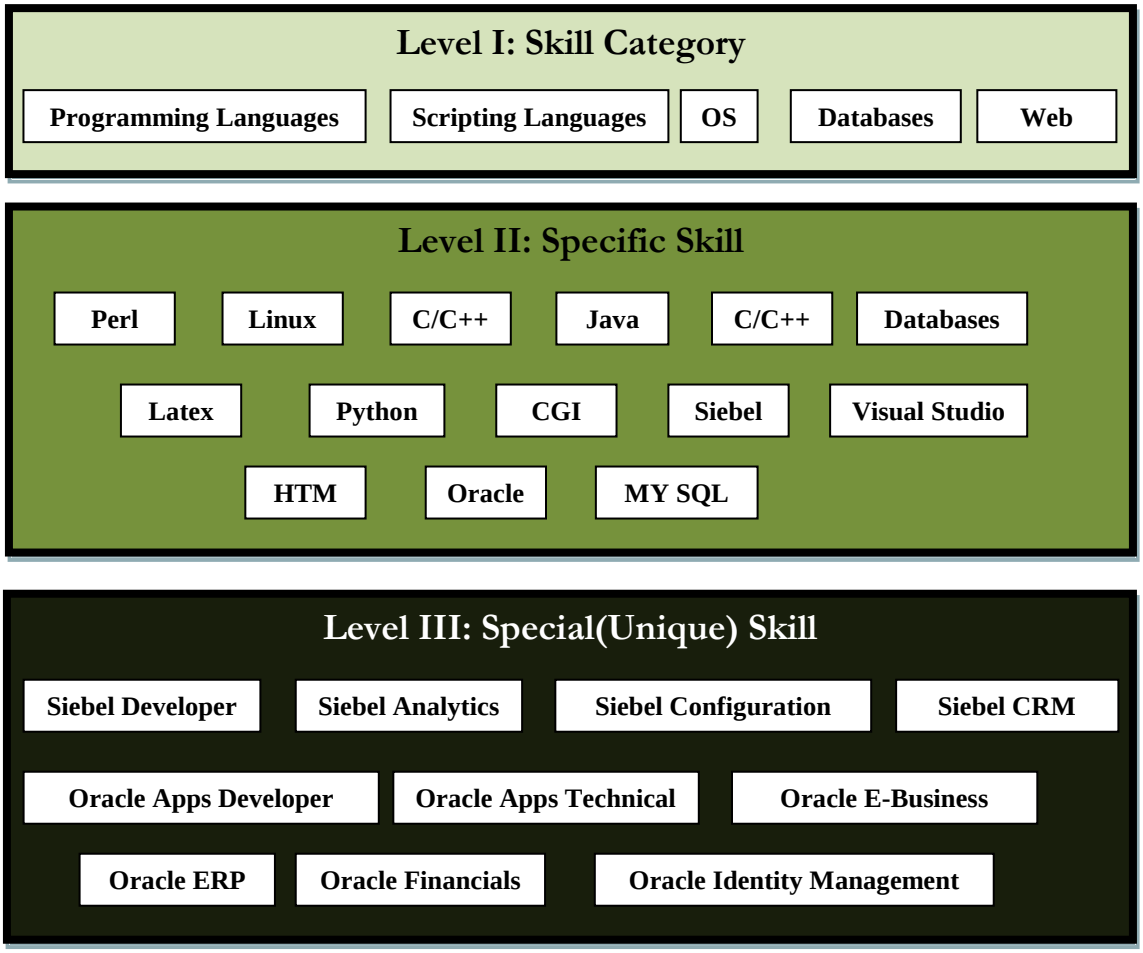


Figure 6: Hierarchical Prototype of three levels of resume skill set

For each feature in S_1, a skill category M is mapped. Similarly for each feature in S_2 is mapped to feature in S_1 and then the corresponding M.

Where S_1 and S_2 are defined as below:

$$S_1 = \mathbf{U}^{n}_{i=1} f_1 (r_i)$$

$$S_2 = \mathbf{U}^{n}_{i=1} f_2 (r_i)$$

For the number of occurrences for each of the skill categories, specific skill and unique skills are counted in each résumé.

This categorization of résumé helps to streamline the initial screening and final résumé selection process. This proposed algorithm is based on the weighted mean of the skills, experience, and

Specific Skills and Corresponding Skill Category

Resume No	Specific Skill	Skill Category	Specific Skill Count
R1	Siebel	Database	2
R1	Oracle	Database	2
R1	Java	Programming	3
R2	Oracle	Database	3
R3	Oracle	Database	1
R2	Java	Programming	2
R8	Oracle	Database	4
R4	Java	Programming	6
R5	Java	Programming	2

Resume No	Skill Category	Count
R1	Database	4
R1	Programming	3
R2	Database	3
R2	Programming	2
R3	Database	1
R4	Programming	6
R5	Programming	2
R8	Database	4

Unique (Special) Skills and Corresponding Skill Category

Resume No	Special Skill	Specific Skill	Skill Category	Count
R1	Siebel Developer	Siebel	Database	1
R2	Oracle Apps Developer	Oracle	Database	2
R2	Siebel CRM	Siebel	Database	1
R2	Oracle Data Operator	Oracle	Database	1
R8	Siebel Test Engineer	Siebel	Database	3
R3	Oracle	Oracle	Database	1

Resume No	Skill Category	Count
R1	Database	1
R2	Database	4
R3	Database	1
R8	Database	3

Figure 7: Examples of Skill Category, Specific Skills and Unique Skills

education in each résumé.

A candidate, who has worked on various professional and educational projects with a particular skill, would have the skill appear frequently in the résumé. We take leverage of this and that helps in adding more value to résumé selection by giving more importance to the candidates who have worked more on a particular technology.

Special skills, specific job profiles, and job titles are extracted for more accurate résumé selection with specific job profiles and job titles. However, we will restrict ourselves on specific skill searches in this project.

Step 2: Initial screening

Initial screening as explained below is done to select the résumés that match each search criteria individually.

a. Select résumés matching the education criteria in the requirement.

   Résumés with particular degree requirements are selected. Each degree obtained is given a score of 1. For example, a candidate holding a Bachelor's and double Master's degree will be scored as Bachelors degree (Score=1) + Double Master's (Score=2). Total education score = 1+2=3. Any score greater than maximum score allowed is treated as achieving maximum score. A maximum education score for this example would be 2 (1 Bachelor's degree + 1 Master's degree). The rèsumès that do not match the education criteria are given the score 0.

b. Select résumés matching required skills.

Check for résumés containing the required set of skills. If the desired set of skills is 3, the maximum score a résumé can get is 3 – one for each skill. If only one of the skills is present in a résumé then the résumé gets the skill score of 1. If none of the skills are available then résumé is 0.

c. Select résumés matching the experience level.

If experience of candidate is an exact match of the job requirement résumé is scored as 2. Otherwise if experience lies within the specified range (Maximum and Minimum) – the résumé is scored 1. The rest of résumé should be given the score 0.

The maximum score is calculated by considering education, experience and skills scoring criteria.

For example, $Education_{Maximum}=2$, $Skills_{Maximum}=3$ and $Experience_{Maximum}=2$ ➔ $Maximum_{[Education, Skills, Experience]} = 3$.

To match all the three scores on equal platform, we normalize the maximum score.

$$Maximum_{[Education=2, Skills=3, Experience=2]} = 3.$$

$$Normalization\ Matrix, N = [Education_{Maximum} *((Maximum)/ Education_{Maximum}),\ Skills_{Maximum} *((Maximum)/ Skills_{Maximum}),\ Experience_{Maximum} *((Maximum)/ Experience_{Maximum})]$$

For the example above, N = [2*(3/2), 3*(1), 2*(3/2)]. For each résumé, multiply all the scores of education by 3/2, skill scores by 1 and education score by 3/2. The sum of all three résumé criteria scores is denoted as Initial Screening score (I_s). A threshold to select the appropriate résumés for next level is computed based on I_s . The résumés with I_s > threshold are selected for final ranking.

Threshold = 50% of SUM [total score] / total number of résumés

Step 3: Final Résumé selection

The final résumé scoring is done to analyze the experience and education more closely from the selected set.

a. Education score is calculated using the exponential calculation scheme where each candidate profile is given a score based on obtained level of educational degree. Each obtained degree is given score as follows:- for Bachelor's (B) as 10, Master's (M) as 100 and PhD (P) as 1000. These individual scores are then added to obtain total education score.

Example 1: Education Score = 1B + 1M+1P = 10 + 100+ 1000 =1110

Example 2: Education Score = 2B + 2M = 2*10 + 2*100 = 220

This scoring system would ensure that the candidates with highest degree requirement are ranked higher in the list.

b. Profiles of candidates are checked for the required job titles held by the person during professional experience. The job titles similar to each other are treated as canonical names and are grouped together.

c. Each résumé is parsed for required skills. Count the number of occurrences of each skill appearing in the résumé. Each skill in the job requirement is given an appropriate weight out of 100 and the sum total of all weights equals 100. A weighted mean of the required skills present in each résumé is calculated and the computed score is called *skill score*. Multiply the skills score with relevant years of experience to obtain *skills-experience score*. The *skills-experience score* will provide the relevant working experience of the candidate with those skills.

d. The *skills-experience score* and *education score* are normalized to bring them to the same scale. The maximum value of *skill-experience score* from the set of résumés is determined. Similarly, maximum value of *education score* is found. The maximum values of the *skill-experience* and *education scores* are considered for normalizing both scores. These normalized scores are then added together to obtain the "*Final Scores*."

e. K-means clustering algorithm is applied on relevant years of experience of a candidate. *K means* divides the working set in to two clusters with minimum and maximum experience values from the set of résumés. This clustering helps in obtaining one cluster that contains résumés with minimum experience or close to minimum experience and another cluster that contains résumés with maximum experience.

f. The top scored résumés above the threshold value are selected from each cluster. A ranking based on the final scores obtained in step (d) is done to prioritize résumés in both the sets.

One of the two sets is selected and is passed to next level for highlighting the specialty of each résumé in the set.

Step 4: Display résumé set based on "Degree of Uniqueness"

Even though the most appropriate résumés are ranked in each cluster, we still need to highlight unique characteristics of each résumé. This process identifies "uniqueness" of the résumés based on close scores. This will help recruiter's decision making process as they have skill featured highlighted for each résumé.

a. We use a *naïve approach* to further group the résumés in clusters. In this section we score the skills from each résumé. Each résumé has set of characteristics that are shared by few other résumés and some special skills exceptional to them called the "**Uniqueness of Résumé".**

b. The uniqueness of each special skill lies between 0 and 1. If the skill is common to all résumés, the value is 0, if it's unique to a single résumé, the value is 1, and else it is calculated by the formula [1-no. of résumés containing that feature/ total number of features].

c. Uniqueness of skills (U_r): Let R_s be set of the selected résumés from a cluster. Each résumé in this cluster possess skills as its characteristic "$f((R_s)_i)$". Each characteristic from a résumé

is denoted by s_j where $0 \leq j \leq |R_s|$ and $n(s_j)$ is the count of résumés to which the characteristic skill belongs. This can be viewed as a multi-set <résumé, skill>.

$$\text{Uniqueness of skill} = \begin{cases} 1 & \text{if } n(s_j) = 1 \\ 1-(n(s_j)/|R|) & \text{otherwise} \end{cases}$$

(where R is the set of 'n' similar resume)

d. The set of résumés is now categorized based on the uniqueness of skills in the set *Rs*. The skills are distributed into three level structures – Level-I, Level-II, and Level-III. The résumé's characteristic skills which are common to all the résumés and have the $U_r=0$ are categorized into Level-I. Level-II contains the skills between $0 < U_r < 1$ and Level-III contains the character skills which have $U_r=1$.

e. The résumés are clustered together based on the Level-II characteristics. The algorithm used to obtain the clusters is explained as follows: Select the first résumé and assign it to Cluster1. The character skills of the first résumé are also assigned to Cluster1. Select the next résumé and compare the character skills with all the clusters and calculate the similarity. A similarity threshold (ST) is calculated by taking the average number of characteristic skills in R_s and eliminating the common skills. If the calculated similarity of the considered résumé is less than ST, then a new cluster is created. The process is repeated to include all the résumés in the selected résumé cluster.

f. The Level-I characteristic skills are listed in common. The Level-II characteristic skills are common for a set of résumés. The characteristic skills and résumés to which those skills belong to are listed along with unique skills of each résumé.

This method of displaying the résumés in the clusters can be extended for the résumé titles, specific job profiles, education or other specific searches.

In Table 1, page 34, the recruiters can review the prospective selected candidates' summary with skills, unique skills, and job titles. The results can be modified to include the criteria of interest. The ranks along with the résumé names give more information about the résumé, thus stating that the particular résumé as a higher precedence compared to the others in the set.

Table 1: Table explains the reduced number of résumés with unique skills and other features.

Level-I: Common Characteristics – SQL, MS-Office			
Technology	Résumé ID	Unique Skills	Required Title
Java, Oracle	1(Rank-10)	Cobol	
	3(Rank-2)	MS-Access	
	9(Rank-5)	Sybase	
J2EE	2(Rank-7)	Cobol	
	4(Rank-3), 8(Rank-9)	Perl, Siebel, Server side scripting	Siebel Development and Test Engineer
	7(Rank-4)	CGI	
Oracle, Perl, Siebel	5(Rank-6)	C/C++	
	6 (Rank-1)	Siebel	Siebel Developer
Java, PHP, JavaScript	11(Rank-15), 16(Rank-17), 19(Rank-12)	Cobol, MS-Access, Android	
	14(Rank-13), 12(Rank-11)	Sybase	
	17(Rank-18)	Cobol	
	18(Rank-19)	Siebel	Siebel Developer
	17(Rank-18)	CGI, Siebel	
	18(Rank-19)	Server side scripting	
Oracle, Siebel	20(Rank-20)	C/C++, Mobile Programming	
	13 (Rank-16)	Sybase	

6. EXPERIMENTS AND RESULTS

In this section, we discuss and analyze the experimental results. To evaluate the performance, we have applied the proposed framework on a real world data-set of résumés.

6.1 DATASET DESCRIPTION

A dataset of 1950 résumés were collected from various sources and annotated so that the results obtained could be validated. The résumé set of 500 was used for training the system. The data was collected from following sources:

- Search Engine: Google, Yahoo, Bing
- Soople.com
- Students and colleagues from Computer, Electrical, Civil Engineering and others.
- Textanalysis.com
- University websites.

The obtained dataset contained profiles from different technical backgrounds, with experience ranging from recent graduates to 30 years. The dataset has profiles from Software engineers, IT professionals, Electrical engineers, Structural engineers, Software testing professionals, Business Analysts, PhD holders, Graduate students, undergraduate students' et all.

All the résumés obtained from various sources were converted to text format for processing. The count of indexes of the obtained document formats is stated below.

Table 2: Résumé dataset formats

Résumé Format	Count
.doc format	661
.txt format	483
.docx format	46
.pdf format	568
.html	192

6.2 EXPERIMENT SETUP

All résumés were converted to text format for processing. The analyzer called “TextAnalysis” was used for information extraction and the generation of separate files for each section of a résumé. Once the dataset is ready, we read the data into the database. We extracted the personal, educational, experience and skill information as required.

A predefined database was used for:

- Resolving canonical names.
- Skill Category
- Specific Skills
- Job Titles
- Education degrees

6.3 ANALYSIS

We define “Test Data Ratio” and “Actual Data Ratio”. These parameters indicate the authenticity which we obtain using our algorithm on the training data set.

$$\text{Test Data Ratio} = \frac{\text{Total number of relevant resumes}}{\text{Total number of resumes in the data set}}$$

$$\text{Actual Data Ratio} = \frac{\text{Total number of actual relevant resumes}}{\text{Total number of resumes in the data set}}$$

We used a set of 500 résumés as a training set. The "Test Data Ratio" vs. "Actual Data Ratio" for 3 different job requirements is summarized in the graph below. As the graph indicates, there is about 10% of an error rate by the manual and the actual data filtering of résumés for a particular job requirement.

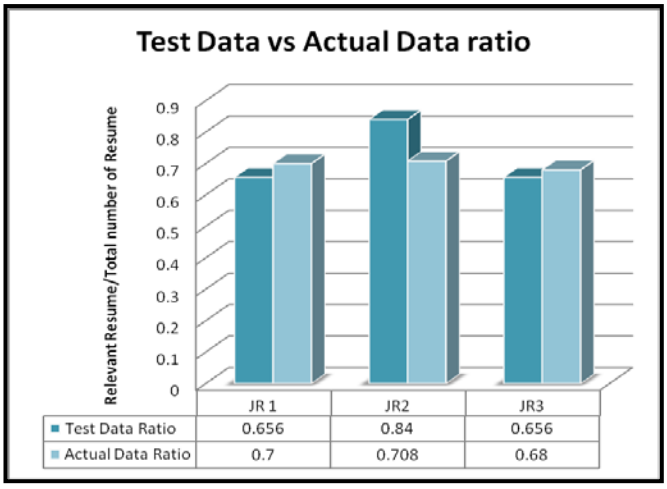


Figure 8: Graph of Test Data vs. Actual Data ratio

For analyzing the framework, all the résumés are divided into sections of education, experience, skills and personal information. The list of skill categories are as follows:

Table 3: Skill Category for selecting Résumé

Skill Category	
programming language	other tools
scripting languages	compiler tools
Libraries	mobile platforms
database technologies	middleware technologies
operating systems	Microsoft Office
testing tools	Mainframes
networking	software tools
web technologies	server side scripting
documentation	open source tools
assembly languages	frameworks and content management systems

The categories presented in the dataset under consideration are ‘operating systems,’ ‘web technologies,’ ‘system programming,’ ’mobile technologies,’ ‘testing,’ ‘database technologies,’ ‘middleware technologies,’ ‘networking,’ ‘complier tools,’ ‘Microsoft office,’ ‘software tools,’ ‘mainframes,’ ‘operating systems’ and ‘server side scripting.’ There are 50 specific skills present in the dataset.

The performance metric is defined as reduction factor (rf) that measures the reduction in the number of résumés under consideration. Let ‘n’ be the total number of résumés that match the criteria using filters and ‘n1’ denote the number of résumés obtained after final processing.

$$rf_{(\text{resume count})} = 1 - \frac{n1}{n}$$

The graph in Figure 9, page 39, provides reduction factor statistics on a training dataset when tested on five identified job requirements.

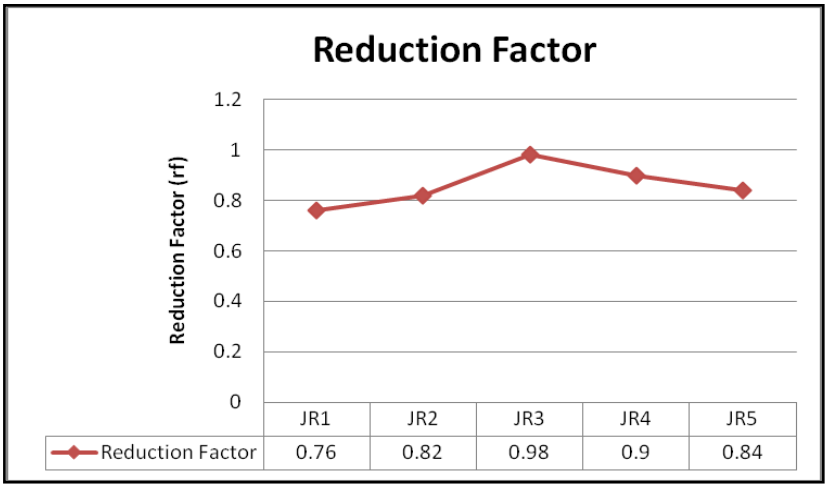


Figure 9: Reduction Factor graph for resumes under consideration

Similarly we calculate the reduction factor of the total number of skills the recruiter needs to consider. Let 'F' denote the total number of skills for all the résumés and F(i) denote the number of features in Level-i and L is 3 denoting number of levels (Level-I, Level-II, Level-III).

$$rf_{(Skills)} = 1 - \frac{\sum_{i=1}^{L} F(i)}{F}$$

The calculated reduction factor formula (above) demonstrating Table 4, page 40, compares the results from *naïve algorithm* [5] and the *proposed algorithm*. Clustering in *naïve algorithm* is done on similar résumés without considering any clustering criteria [5]. It reduces the number of skills to be considered by recruiters. This algorithm fails to toss out résumés that do not match certain job requirements and clusters all the résumés under consideration. On the contrary, the *proposed algorithm* considers all the factors and clusters relevant résumés only. Selected résumés in the cluster are ranked and can be easily compared against one another as they match

determined job requirements. Ranking résumés in this way helps to identify relevant résumés among a given set. This approach can be called "Relevant résumé skill clustering." The 'rf' for "Relevant résumé skill clustering," is 80% of the original dataset, whereas 'rf' for naïve clustering is 74%. There is a 6% increase in the 'rf' value when 'Relevant résumés' are considered compared to 'Similar résumés' using the same dataset.

Table 4: Reduction Factor Values for specific skills

Feature Type	|F|	$\sum_{i=1}^{L} F(i)$	rf
Specific Skills (Using Naïve Clustering [5])	800	**210**	**0.74**
Specific Skills (Using Proposed Algorithm)	554	**91**	**0.82**

Below is the 'rf' graph comparing the results of '*naïve clustering*' and '*relevant résumé skill cluster*' using five job requirements. To obtain the results, we executed both the algorithms on

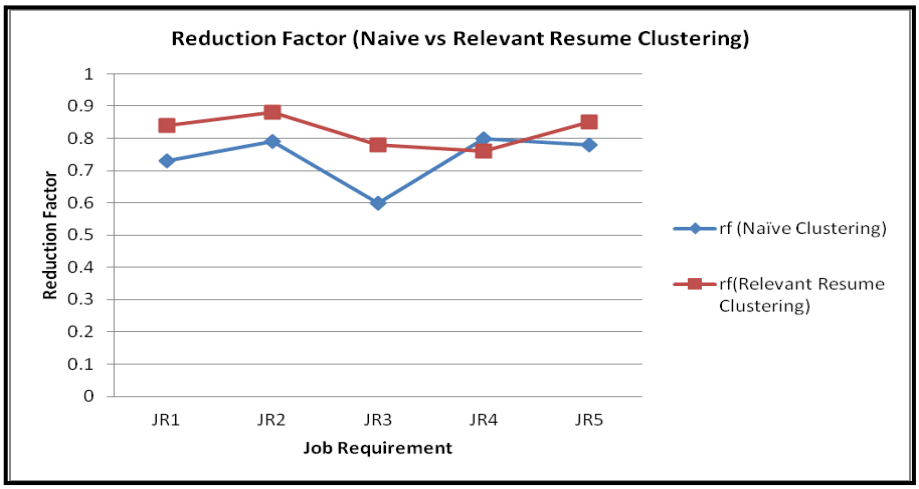


Figure 10: Naïve Clustering vs. Relevant Resume Clustering

the data set and calculated the reduction factor. Results clearly indicate that clustering relevant résumés would be more efficient and also useful for recruiters than just clustering similar résumés.

7. CONCLUSION

There are problems in résumé processing and the selection of appropriate résumés from an assortment of defined résumés. We have made an attempt to pull out selected résumés by choosing them based on defined job requirements and then highlighting their unique features. The approach used here uses a scoring technique to select the best résumé. The skill categories, specific skills, and unique skills of a résumé are considered to determine the uniqueness of a résumé. This helps recruiters speed read through a set of résumés and their specialties in order to decide on prospective candidate. Résumé ranking helps to provide an appropriate résumé based on the criteria matched.

The technique used considers a set of résumés from those based on calculated scores. Each résumé is eliminated based on a low score. When compared with the results of current résumé filters, certain résumés are not selected based on keyword search. This avoids the possibility of rejection of a résumé based on the order of filtering criteria during the selection process.

The algorithm that highlights the unique features of selected résumés would help recruiters obtain the essence of each résumé. The process helps in easy résumé selection. The results obtained using the, “Relevant Résumé Skills,” for displaying unique skills demonstrated better results compared to “Similar Résumé” in naïve algorithm that does not consider job requirements [5] by 6%. The approach proposed is the foundation and can pave way toward solving the problem of résumé selection. Additional research work is required to address the problem of specific résumé selection and extraction.

8. FUTURE WORK

The proposed approach has a potential for improvement and can include other features that appear in a résumé such as certifications, co-curricular activities, interests, etc. The weighted scores for skills can also be extended to an initial screening scoring method. For example, one can calculate the maximum number of skills under consideration and give a weighted unit for each skill instead of giving a unit value. This would help rank résumés according to the desired weighted unit.

Investigations can extract the context of a job profile selected by calculating the distance between words. For example: résumés with keywords as ‘Sybase,’ ‘testing,’ and ‘testing on Sybase’ should be treated in a different way and a recruiter looking for a Sybase tester could be provided with relevant résumé and not ones with keywords such as ‘sybase’ and ‘testing’ in same or different sections.

9. REFERENCES

[1] Jonathan Medema "Reliable Normalization in Résumé Information Extraction," Published in 2008 http://igitur-archive.library.uu.nl/student-theses/2009-0211-202406/UUindex.html

[2] Roger E. Bohn and James E. Short, "How Much Information? 2009," http://hmi.ucsd.edu/pdf/HMI_2009_ConsumerReport_Dec9_2009.pdf

[3] Gang Guan, Kun Yu and Ming Zhou "Résumé Information Extraction with Cascaded Hybrid Model" ACM pages 499–506, Ann Arbor, June 2005. <http://acl.ldc.upenn.edu/P/P05/P05-1062.pdf>

[4] B. L. Hawkins, J. A. Rudy and W. H. Wallace "Recruiting, Retaining, and Reskilling Campus IT Professionals. Technology Everywhere: A Campus Agenda for Educating and Managing Workers in the Digital Age". Dolan, A. F. 2004. Jossey-Bass: 75--91.

[5] Sumit Maheshwari "Mining Special Features to Improve the Performance of Product Selection in E-commerce Environment and Résumé Extraction System" (ICEC 2009) Taipei, Taiwan, August 2009, Published by ACM.

[6] Optimizing Your Résumé for Scanning and Tracking Systems <http://www.montclair.edu/CareerServices/OptimalsScannedrésumés.pdf>

[7] Kun Yu, Gang Guan, and Ming Zhou, "Résumé information extraction with cascaded hybrid model". In ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pages 499–506, Morristown, NJ, USA, 2005.

[8] Sumit Maheshwari, Abhishek Sainani, and P. Krishna Reddy, “An Approach to Extract Special Skills to Improve the Performance of Résumé Selection”. In Proceedings of DNIS. 2010, 256-273.

[9] Abhishek Sainani, “Extracting Special Information to Improve the efficiency of Résumé Selection Process,” June 2011.

[10] Résumé Manager Pro www.sarmsoft.com

[11] Andrew Kantor, “Internet suffering from information overload,” http://www.usatoday.com/tech/columnist/andrewkantor/2007-06-14-internet-organization_N.htm

[12] Xing Yi, James Allan and W. Bruce Croft, “Matching Résumés and Jobs Based on Relevance Models”. In Proceedings, SIGIR 2007.

[13] Payal Chanania, “Use proactive techniques to filter résumés,” The Hindu - Opportunities <http://www.hindu.com/jobs/0708/2007082950020200.htm>

[14] Norm Schneider, “Job Hunters: Résumé Filters May Help or Hinder Your Job Search,” Published on August 23, 2011 Employment http://bizcovering.com/employment/job-hunters-résumé-filters-may-help-or-hinder-your-job-search/#ixzz1fajVNEJ2.

[15] TAI. 2003. VisualText Help. Text Analysis International http://www.textanalysis.com/help/help.htm

[16] Kardi Teknomo, “K-Means Clustering Tutorial,” http://www.croce.ggf.br/dados/K%20mean%20Clustering1.pdf

[17] Daxtra CVX. http://www.daxtra.com/

[18] Sovren Résumé/CV Parser. http://www. sovren.com/

[19] ALEX Résumé Parsing. http://www.hireability.com/alex/

[20] Akken Stuffing. http://www.akken.com/.

[21] RésuméGrabber Suite. http://www.egrabber.com/résumégrabbersuite/.

[22] F. Ciravegna and A. Lavelli. “Learning Pinocchio: adaptive information extraction for real world applications”. Nat. Lang. Eng., 10(2):145{165, 2004}.

[23] Z. Jiang, C. Zhang, B. Xiao and Z. Lin “Research and Implementation of Intelligent Chinese Résumé Parsing,” Proceedings of the 2009 WRI international Conference on Communications and Mobile Computing, vol. 3, 2009.

[24] Vinayak Borkar, Kaustubh Deshmukh and Sunita Sarawagi, “Automatic segmentation of text into structured records”. In Proceedings of SIGMOD 2001 International conference.

[25] Ertug Karamatli and Selim Akyokus. “Résumé information extraction with named entity clustering based on relationships”. INISTA 2010, 2010.

[26] TalentDrive. www.talentdrive.com/

[27] Résumé Dragon résumédragon.com/

[28] Jakub Piskorski, Marek Kowalkiewicz, and Tomasz Kaczmarek “Information extraction” from cv. pages 185–192, 2005.

[29] Nakayama, M and Sutcliffe, N.G. (2007). “Perspective-Driven IT Talent Acquisition”. Proceedings of SIGMIS-CPR’07, April 19–21, 2007, St. Louis, Missouri, USA. ACM 978-1-59593-641-7/07/0004, 2007.

[30] A.R. Ahlan, Y. Arshad M.A. Suhaimi, and H. Hussin. “The future skill sets expectations of IT graduates in Malaysia IT outsourcing industry,” presented at the 7th WSEAS International Conference on E-Activities, Cairo, Egypt, Dec 29-31, 2008, Paper 605-256.

[31] Maheshwaran Janarthanan, Devesh Kumar Singh “Web Mining Résumé”.

[32] Brian Hawkins, Julia A. Rudy, and William H. Wallace, Jr., eds., “Technology Everywhere: A Campus Agenda for Educating and Managing Workers in the Digital Age”, EDUCAUSE Leadership Strategies Series, vol. 6 (San Francisco: Jossey-Bass, 2002), http://www.educause.edu/LibraryDetailPage/666?ID=PUB7006.

[33] Allen, Claudia. “The Career Center: Becoming Paperless is a Bonus”. Journal of Career Planning & Employment, v57 n3 p25,29-31 Spring 1997

10. APPENDIX

Consider a job following job requirement:

Required Skills:

- 6 years of experience as Sybase Administrator with Masters in Computer Science.
- Should be a very good team player.
- Should be able to support high volume trading application
- Strong troubleshooting and debugging skills.
- Good understanding and working knowledge of Sybase database architecture.
- Very strong in Sybase T-SQL programming like writing stored procedures, cursors, triggers and performance tuning
- Strong in using Sybase and MSSQL client utilities like BCP, Defncopy etc.
- Very strong in PERL and UNIX Shell scripting.
- Very good in DBA tasks like installing new servers, upgrades, tuning of the Sybase servers, creating new databases, backup/recovery etc.
- Equity trading experience is a plus.

Figure 11: Job requirement

Summary of the job requirement:

Experience: 6 years (Assume the employer is fine with Experience range from 5years to 8 years)

Title: Sybase Administrator, DBA

Education: Master's in Computer Science

Skills: Database (Sybase, T-SQL, MSSQL), Scripting (Perl, UNIX Shell Scripting)

Special skills: Equity trading

To process the above job requirement, and determine the prospective candidate profiles, we would use the "Relevant Résumé Skill Clustering approach."

Step 1: Initial Filtering

Table 5: Initial Screening Data

Résumé Name	Skill	Experience	Education	Total		Résumé Name	Skill	Experience	Education	Total
R5	5	5	0	10		R27	0	0	5	5
R12	4	2.5	5	11.5		R28	0	0	5	5
R50	3	2.5	5	10.5		R40	0	0	5	5
R13	2	2.5	5	9.5		R45	0	0	5	5
R14	2	2.5	5	9.5		R18	5	0	2.5	7.5
R38	2	2.5	5	9.5		R7	4	0	2.5	6.5
R26	1	2.5	5	8.5		R17	4	0	2.5	6.5
R44	4	2.5	2.5	9		R30	4	0	2.5	6.5
R49	3	2.5	2.5	8		R37	3	0	2.5	5.5
R19	2	2.5	2.5	7		R48	3	0	2.5	5.5
R32	1	2.5	2.5	6		R6	2	0	2.5	4.5
R2	2	2.5	0	4.5		R9	2	0	2.5	4.5
R21	5	0	5	10		R25	2	0	2.5	4.5
R35	5	0	5	10		R31	2	0	2.5	4.5
R41	4	0	5	9		R42	2	0	2.5	4.5
R8	3	0	5	8		R47	2	0	2.5	4.5
R15	3	0	5	8		R1	1	0	2.5	3.5
R23	3	0	5	8		R24	1	0	2.5	3.5
R34	3	0	5	8		R36	1	0	2.5	3.5
R39	2	0	5	7		R43	1	0	2.5	3.5
R11	1	0	5	6		R22	0	0	2.5	2.5
R16	1	0	5	6		R3	4	0	0	4
R29	1	0	5	6		R20	4	0	0	4
R46	1	0	5	6		R4	3	0	0	3
R33	1	0	0	1		R10	1	0	0	1

Calculate threshold: 70% of [Sum of Total/Total Number of résumé]

(Assume that we want to set the threshold as 70% to have more precise résumé search.)

= 70% [310/50] = 4.34

All the résumés with total less than 4.34 will not be considered in the second step. The resultant set for the next pass is in the table below:

Table 6: Data obtained after Initial Screening and threshold

Résumé Name	Skill	Experience	Education	Total		Résumé Name	Skill	Experience	Education	Total
R5	5	5	0	10		R46	1	0	5	6
R12	4	2.5	5	11.5		R33	1	0	0	1
R50	3	2.5	5	10.5		R27	0	0	5	5
R13	2	2.5	5	9.5		R28	0	0	5	5
R14	2	2.5	5	9.5		R40	0	0	5	5
R38	2	2.5	5	9.5		R45	0	0	5	5
R26	1	2.5	5	8.5		R18	5	0	2.5	7.5
R44	4	2.5	2.5	9		R7	4	0	2.5	6.5
R49	3	2.5	2.5	8		R17	4	0	2.5	6.5
R19	2	2.5	2.5	7		R30	4	0	2.5	6.5
R32	1	2.5	2.5	6		R37	3	0	2.5	5.5
R2	2	2.5	0	4.5		R48	3	0	2.5	5.5
R21	5	0	5	10		R6	2	0	2.5	4.5
R35	5	0	5	10		R9	2	0	2.5	4.5
R41	4	0	5	9		R25	2	0	2.5	4.5
R8	3	0	5	8		R31	2	0	2.5	4.5
R15	3	0	5	8		R42	2	0	2.5	4.5
R23	3	0	5	8		R47	2	0	2.5	4.5
R34	3	0	5	8						
R39	2	0	5	7						
R11	1	0	5	6						
R16	1	0	5	6						
R29	1	0	5	6						

Step 2: Final résumé selection

To calculate the résumé relevant index, we need to score the education and experience-skill.

Score for Master's =100, Score for Bachelor's=10, Maximum score=110.

Step 3: Consider the résumés with “Sybase Database Administrator” as the job title and calculate the experience as Sybase administrator.

Calculate the total experience as the Database administrator (canonical name)-“Sybase Administrator”.

Scan the résumé to determine the count of skills present in each. For example: in this example, Sybase is more important and so the weight can be as follows: Sybase (50), T-SQL(15), MSSQL(15), Perl(10), and Unix Shell Scripting(10). Weighted mean is calculated to find skill_score.

This skill_score*experience as Database Administrator + Experience as “Sybase Database Administrator” provides the relevant working experience.

Normalize the skill experience and education score.

Table 7: Data after Normalization

Résumé Name	Skill-Experience	Education	Normalize	Education	Total		Résumé Name	Skill-Experience	Education	Normalize	Education	Total
R46	3.69	110	90.2	110	200.2		R32	1.35	20	33	20	53
R38	3.4	110	83.111	110	193.1111		R28	1.66	10	40.6	10	50.5778
R26	2.38	110	58.178	110	168.1778		R14	1.35	10	33	10	43
R13	2.13	110	52.067	110	162.0667		R2	1.35	10	33	10	43
R44	2.13	110	52.067	110	162.0667		R21	1.35	10	33	10	43
R49	2.13	110	52.067	110	162.0667		R35	1.35	10	33	10	43
R50	2.13	100	52.067	100	152.0667		R31	1	10	24.4	10	34.4444

R16	1.59	100	38.867	100	138.8667		R5	0.77	10	18.8	10	28.8222
R18	0.98	110	23.956	110	133.9556		R9	1.02	0	24.9	0	24.9333
R41	1.35	100	33	100	133		R8	0	20	0	20	20
R19	4.5	20	110	20	130		R42	0.38	10	9.29	10	19.2889
R25	4.13	10	100.96	10	110.9556		R7	0.3	10	7.33	10	17.3333
R34	0	110	0	110	110		R11	0.18	10	4.4	10	14.4
R29	0	110	0	110	110		R15	0.03	10	0.73	10	10.7333
R23	0.16	100	3.9111	100	103.9111		R39	0	10	0	10	10
R6	0	100	0	100	100		R33	0	10	0	10	10
R37	2.67	10	65.267	10	75.26667		R27	0	10	0	10	10
R17	2.04	10	49.867	10	59.86667		R40	0	10	0	10	10
R12	2.38	0	58.178	0	58.17778		R30	0	10	0	10	10
R48	1.77	10	43.267	10	53.26667		R47	0	10	0	10	10

Apply the K-means clustering on the experience by choosing the 2 clusters with minimum and maximum values and calculate the threshold to find the accept values.

Table 8: Relevant Résumés

Résumé Name	Total	Rank
R46	200.2	1
R38	193.1111	2
R26	168.1778	3
R13	162.0667	4
R44	162.0667	5
R49	162.0667	6
R5	152.0667	7
R16	138.8667	8
R18	133.9556	9
R41	133	10
R19	130	11

R25	110.9556	12
R37	75.26667	13
R17	59.86667	14
R12	58.17778	15

Step 4: Display résumé set based on Uniqueness of résumé.

The selected relevant résumés are arranged based on the "Degree of Uniqueness" for each skill present in the cluster of résumé. This view can be used by recruiters to get the idea of the skills present in the selected résumé set. The numbers in front of the Résumé name indicate the ranks of the résumé calculated based on the scores.

Table 9: Unique Features of a Résumé

Common Clusters : C/C++, SQL			
Specific Skills	Résumé Name	Unique Skill	Title
Sybase, Oracle	R46 (1)		
	R38 (2)	Perl	
	R26 (3)	J2EE	
	R13 (4)	T-SQL	
JavaScript, Sybase, Unix	R44 (5), R49 (6), R5 (6)	Oracle	Database Sybase Administrator
	R25 (12), R37 (14)		
	R16 (7)	Matlab	
	R18 (8)	PHP	
Perl, Java	R41 (9)	Sybase	Database Sybase Administrator
	R19 (10)	MySQL	
Unix Shell Scripting, Perl	R17 (15)	Sybase	
	R12 (15)	T-SQL	