Delivering Value from Resume Repositories

Executive Summary

Repositories of resumes, of both external candidates and current employees, contain valuable information. We identified several use-cases where information and insights obtained from resumes can improve the quality of work as well as satisfaction levels for employees, in addition to improvements in HR processes. We have briefly reviewed several technologies such as natural language processing, text-mining, data-mining and information retrieval, which are mature enough and can be applied to processing and analysis of resume repositories. We have suggested that it is possible to build right set of resume processing and analysis tools, respecting the candidates’ privacy and complying to the applicable data privacy laws, which are closely aligned to each of the identified use-case scenarios and deliver high-quality information and insights required there. One way forward would be to build a central system, which may be called ResumeCenter, which delivers all services related to processing and analysis of resumes to various users and systems across an organization.

About the Authors

Girish Keshav Palshikar is a Principal Scientist and leads the Machine Learning R&D group in Systems Research Lab (SRL), an innovation center in Tata Consultancy Service Limited, Pune, India. His research interests are in machine learning, data mining, text mining and theoretical computer science. He has more than 60 publications in international journals and conferences.

Rajiv Srivastava leads the iWork program in SRL aimed at building tools and systems related to HR analytics. Rajiv is building ResumeCenter, a single-window system to deliver information and insights from processing and analysis of resume repositories. Many parts of ResumeCenter are already functional and deployed for various users in TCS HR.
# Table of Contents

1. Introduction ........................................................................................................................................ 3  
2. Structure of a Typical Resume ............................................................................................................. 3  
3. Resume Repositories .......................................................................................................................... 4  
4. Scenarios for Using Resumes ............................................................................................................. 5  
   4.1 Talent Acquisition ......................................................................................................................... 5  
   4.2 Talent Management and Project Leadership ............................................................................... 6  
   4.3 Senior HR Management .............................................................................................................. 6  
5. Challenges in Resume Processing ...................................................................................................... 6  
6. Systems for Resume Processing ......................................................................................................... 7  
   6.1 Resume Processing in Most Organizations ................................................................................. 7  
   6.2 Existing Solutions for Resume Processing ................................................................................... 7  
7. Technologies for Resume Processing ................................................................................................. 8  
   7.1 Information Extraction ................................................................................................................. 8  
   7.2 Information Retrieval .................................................................................................................. 9  
   7.3 Data and Text Mining for Improving Efficiency of HR Tasks .................................................... 9  
8. The ResumeCenter Solution .............................................................................................................. 10  
9. Conclusions ....................................................................................................................................... 11  
References .................................................................................................................................................. 12
1. Introduction

A **resume** (also called **CV** or **bio-data**) is a vital source of information about the education, skills, experience and expertise of a person. A resume contains a summary of the entire work experience of a person and includes personal details, educational qualifications, job history, projects and tools worked on, trainings, awards, publications, professional memberships and activities etc. Resumes of candidates arrive at an organization from different **sources** (also called **modes** or **channels**) such as postal mails, emails, uploads on organizations career portal, recruitment agencies, job sites, external consultants, campuses, walk-in, employee referrals etc. Most organizations also maintain repository of the resumes of their current employees, which are then used for various purposes such as allocation to projects, interactions with clients etc. Resumes are regularly updated by the authors – candidates or employees – to have favorable first impression on the prospective employer.

Resumes are typically free-form or semi-structured English documents received in text (.TXT), Microsoft Word (.DOC, .DOCX, .RTF), HTML or PDF formats. Resumes may occasionally contain tables or images. There is no standard way of writing a resume and the resumes in any large repository display a bewildering variety of formats, vocabulary, writing styles, type of information contained and writing quality. A resume from a candidate usually has dual (and often conflicting) goals: to provide factual information (personal, academic, professional) and to impress the prospective employer about one’s suitability and competence for a job.

2. Structure of a Typical Resume

As mentioned, there are large variations in the structure of resumes received from outside an organization. The structure of resumes may also depend on the domain; e.g., resumes of IT engineers versus banking professionals. A resume is typically divided into **sections**, some of which are present in almost all resumes and some sections may be optional (Fig. 1).

- Career Objectives
- Employment Summary
- Educational Details
- Personal Information
- Skill Profile
- Experience in detail (covering projects or job descriptions)
- Training and Certifications
- Publications
- Awards and Recognitions

**Figure 1. Sections in a typical resume.**

**Career objectives** section captures the attitude and intent of the candidate. It is a high-level, qualitative description of the services and skilled work that the candidate is eager to render.

**Employment Summary** provides a brief description of the jobs previously done. For each job, typically the description consists of the period, name of company or organization and the highest designation or position held. Usually a tabular format is preferred by candidates. There are resumes, mostly for senior
management candidates, wherein the job description is given in running text capturing details about work domain and duration of roles played.

**Educational Details** consist of degree name, date or year of passing, university and/or college name (sometimes with location), university name (sometimes with location), specialization (or branch) and the class or grade obtained. The class values are textual and grade values are often mentioned as alphabets along with CGPA score. The institute, university name and even the specialization may be missing. Degrees are written in many different ways, as in *Master of Computer Applications, MCA* or *M.C.A.*

**Personal Information** is usually not a single contiguous section. The personal information may appear at various points in the resume; e.g., contact details at the beginning and some other details towards the end. Contact details consist of name, current and permanent postal addresses, mobile numbers, phone number and e-mail. Other personal information consists of date of birth (age), marital status, languages known along with reading, writing and speaking proficiency, passport number along with place of issue and date of validity, visa taken from countries and their validity, father’s name, mother’s name, hobbies, extra-curricular activities etc.

**Skill Profile** covers the candidate’s experience in technical skills, tools and technology, domain-related platforms. For IT resumes, the skills are divided into groups such as programming languages, software tools, methodologies, IDEs, Hardware, operating systems, application packages, frameworks, databases etc. The managerial skills are also covered in the form of processes, methodologies etc.

**Experience** section provides real insight into the job or projects carried out by the candidate. This section not only provides details of the work done like client name and location, domain, duration, skills used, role played but also gives a true picture of the skills learnt and experienced on the job. The experience section is organized as *per job or per project*, typically in reverse chronological order. Usually there is also a textual *description*, which describes the client’s business, brief overview of the system, tasks handled, tools used, methodologies followed, notable achievements or successes etc.

**Training and Certifications** is typically a separate section but may sometimes be clubbed with the Educational Details. Each certification has name of certification, certifying authority and date of certification. Training details typically include these details in addition to the subject of training and location where the training was conducted.

**Publications** section is often included in the resumes of senior candidates or those with research background. Each publication typically mentions authors, title, journal/magazine/conference name, volume/issue and page numbers, year of publication etc.

**Awards and Recognitions** are usually mentioned in a separate section. The awards may be of national or international stature, or may be at university, college and organization level, typically mentioned by the younger candidates.

### 3. Resume Repositories

It is not unusual for an organization to collect hundreds of thousands of resumes, or even millions of resumes in case of large organizations or job portals. It is important to organize these resumes for effective and efficient usage. At the time of registration of candidates, the channels may also obtain and store a limited amount of *structured data*, such as name, address, contact information, skills, work
experience etc. using a simple form. This auxiliary structured data, along with the resume itself, needs to be linked and stored. The candidates also update their resume as well as the associated structured from time to time. In addition, HR executives and other users of resumes may write and store notes, comments or annotations on a resume; e.g., remarks on skills, experience, personal nature etc., which are a valuable part of the resume data repositories (Fig. 2).

Figure 2. Types of annotations on resumes.

In the simplest case, a resume repository can consist of simple folders where the resumes are stored. Alternatively, resumes can be stored in structured databases (such as Oracle), which allow storing of text files as part of the data; additional columns can contain users’ annotations and structured data such as name, date of upload etc. Repositories of employees’ resumes and candidates’ resumes are important reservoirs of information, and can be used in various ways by HR executives (in Talent Acquisition, Talent Management), managers and project leaders in an organization. Managing resume repositories can be a daunting task, where special issues like identifying duplicate and stale resumes need to be handled.

4. Scenarios for Using Resumes

Some practical use-case scenarios where information from resumes is needed are as follows:

4.1 Talent Acquisition

- **Screening (performing validation checks)** to detect and reject ineligible, irrelevant or suspect resumes; e.g., eligibility rules, duplicate resumes, gaps/errors in data provided, criminal misrepresentation such fake degree/university or fake experience descriptions.
- **Resume Analysis (Tagging)** to extract and annotate special qualities, attributes, information etc. from a resume (Fig. 2).
- **Short-list** the candidates best matching with the skills and experience requirements for specific positions in a group, project, account or location.
- **Update internal HR data** by extracting relevant information from employee resumes; e.g., update employee training/skills along with proficiency levels
- **Routing resumes to right persons** by matching resume to domains and available job descriptions
- **Interviewing the Candidates**, where information from resumes is used extensively.
4.2 Talent Management and Project Leadership

- **Create/update skill profiles** of employees working in existing teams (project, account, location, business unit) with respect to competencies and proficiencies in various kinds of skills (technical, quality, methodological, managerial) and upload these profiles into internal HR systems.
- **Identify skill gaps** of a team (project, account, location, business unit etc.), particularly with respect to existing and upcoming projects.
- **Recommend training programs** to employees to upgrade skills and for career planning.
- **Form optimal teams** by selecting right members best matched for a set of projects/roles.
- **Identify experts** for specific technical areas / domains for proposals or client interactions.

4.3 Senior HR Management

- **Create job market intelligence** (trends/patterns that characterize available talent pool) that can help in optimal recruitment planning.
- **Create technology profiles** (trends/patterns) that can help in optimal skill upgrade planning.
- **Understand factors and root-causes for attrition and build predictive attrition models**.
- **Build indicators for identifying high-quality candidates**: correlate performance of employees / candidates with work history and information in resumes (certifications, awards, academic performance, tier of college, tier of employer, team sizes, roles, responsibilities handled, quality of work etc.).
- **Build indicators for identifying high-stability candidates**: correlate work history (e.g., job changes) of employees with information in resumes (native place, work locations, marital status, number of dependents, education etc.).
- **Improve HR processes**: reduce time and cost of hiring, reduce bench strength, prepare hiring plans for meeting recruitment targets for project pipelines and attrition, handle work-related issues affecting employee satisfaction levels, improve organization-wide competency levels and role profiles, improve quality of people hired, reduce long-term and infant attrition, improve utilization and costs of hiring using various channels, improve productivity of HR executives.

5. Challenges in Resume Processing

Clearly, making an effective use of resumes in a repository is critical for an organization. But several factors inhibit this.

- The number of employee resumes in large enterprises may range from few thousand to few hundred thousand. Number of resumes from candidates who have applied to a large organization may be over a million. This large size means that only the simplest tasks can be done manually.
- The resumes are at best semi-structured English documents, containing no fixed set of sections, no fixed ways of indicating section headers, may contain tables and images, multiple columns etc.
- Resumes may use domain-specific vocabulary (e.g., *Windows DLL on SQL Server*), fragmentary text possibly with abbreviations and punctuation, errors of various kinds (capitalization, punctuation, spelling or grammatical), variations in writing styles, and other formatting and linguistic variations.
- There are no ready-made fixed dictionaries of the entities to be extracted (e.g., skills, designations).
- The flexibility of natural language allows writers to state the same kind of information in many different ways (e.g., *fix, repair, mend*), understanding which may need deep linguistic knowledge.
- Understanding some information in resumes may require domain knowledge (e.g., *costing, cash-flow* and *accounts payable* are related to *finance*).
• There are large variations in the way even purely factual information is stated; e.g., dates may be
written in many ways (e.g., Jan. 31, 2008 or 31-Jan-2008), marks may be written as % or grade
points out of 10.

All this makes it difficult to manually locate and extract information and to compare resumes. HR
executives, managers and project leaders need to manually read resumes and extract different types of
information from them – a time-consuming, error-prone and subjective task, particularly when dealing
with thousands of resumes. Typical side effects of manual processing and analysis of resumes are:

1. Under-utilization of available talent pool
2. Formation of lower quality project teams leading to project failures, employee unease and client
dissatisfaction
3. Higher costs, higher time and lower quality of recruitment
4. Reduced understanding of the organization-wide talent profile, training needs, roles profile etc.
5. Reduced quality of career planning suggestions and counseling help to employees.

Fortunately, natural language processing, text-mining, data-mining and information retrieval
technologies are mature enough and can be applied to processing and analysis of resume repositories. It
is possible to build right set of resume processing and analysis tools, which are closely aligned to each of
the above use-case scenarios and deliver high-quality information and insights required there.

6. Systems for Resume Processing

6.1 Resume Processing in Most Organizations

Most large organizations already have rudimentary and fragmented systems to support maintenance of
resumes and associated information. Some examples are as follows.

• A web-based careers portal may allow candidates to upload resumes, which are stored in a
repository
• Application which allows employees to create and update important structured fields in their
resumes (e.g., the ResumeBuilder system in TCS).
• Internal databases which allows employees to update their skills, competencies and trainings data.

Most such systems are usually not well-integrated with HR processes; e.g., they do not deliver resume-
based information and functionality to appropriate users as and when required.

6.2 Existing Solutions for Resume Processing

Several software organizations and job portals have already realized the importance of a comprehensive
product or solution of processing and analysis of resume repositories. Standard document repository
management solutions (e.g., Documentum) can be adapted and used for managing resume repositories
(e.g., storage, searching). We have already discussed the use of search engines, such as Lucene or
Google Desktop, for searching a resume repository. Popular tools for HR business process and work-flow
management – such as PeopleSoft, OracleApps – usually provide only some basic resume repository
management functionality; e.g. FITS (www.justfits.com) and Resume Repository. There are some tools,
such as Daxtra (www.daxtra.com), ResumeMirror (www.talenttech.com), for resume parsing, which
essentially extract some simple kinds of structured information from resumes and store it in a structured data repository. Some of these tools can be “plugged into” document editors or enterprise solutions like PeopleSoft and Oracle Apps. Recently IBM, India in a press release mentioned a resume mining tool for the low-skilled manual trade/semi-skilled job opportunities for the use of a state employment exchange. The extraction support for the personal details and trade related few entities are supported.

7. Technologies for Resume Processing

Processing a large repository of resumes needs a range of advanced computing technologies. We now briefly review some of these technologies.

7.1 Information Extraction

*Information Extraction (IE)* is a process of extracting structured information from documents containing unstructured natural language text. The extracted information is output to structured data repositories such as database tables or XML files. This structured information can then be effectively searched, disseminated, reused or mined (using data mining techniques) to discover valuable knowledge and insights. The extraction process is often guided by means of extraction patterns (Fig. 3). Accuracy and speed of extraction are two important criteria for designing IE systems. See [1], [2] for overview of the IE technology.

![Diagram of Information Extraction from Resumes](image)

Typically, the information to be extracted from a resume consists of the following type:

(a) **Generic named entities (NE):** names of people, places, organizations, dates, times, emails, URLs, phone numbers, addresses, amounts etc.

(b) **Domain-specific named entities:** e.g., technology skills *(Windows, Java Struts, Oracle 11i)*

(c) **Taxonomic relations:** standard relations between entities, such as IS_KIND_OF, which are not explicitly stated; e.g., *Oracle IS_KIND_OF database.*

(d) **Domain-specific relations:** e.g., a ternary relation between POST, ORGANIZATION and DURATION *(worked as a Project Leader in IBM from Jan. 2008 to Nov. 2010).*

The extracted information is often a “polished” and “standardized” version of the input text.
The crux of IE is the power and ease of use of the way the extraction patterns are defined. Typically, the extraction patterns are defined using a formalism called regular expressions (RE). One can define a set of RE patterns for an entity. The IE engine matches the given patterns with the given input text and extracts the matched text, which corresponds to an occurrence of that entity. As an example, an organization can often be recognized as a sequence of capitalized words ending in a cue word such as Inc., Institute or Company (e.g., Microsoft Inc.). We can see the limitations of such a pattern right away. For example, some organizations may contain other cue words (e.g., Limited), some organizations may contain non-capitalized words (e.g., Royal Bank of Scotland) and some may not contain any known cue word (e.g., IBM or Mitsubishi). While we can keep enhancing the RE pattern to handle some of these failures, clearly such purely lexical patterns are inherently brittle. Humans recognize occurrence of an organization in a sentence using patterns that are much deeper, based on the ways an organization is used in a sentence. The RE pattern can make use of such syntactic or semantic knowledge.

A powerful alternative solution is to use supervised machine learning techniques (e.g., decision trees or support vector machines) to automatically create patterns from a large corpus of labeled examples of the entity. Or one can use unsupervised machine learning techniques to automatically create a list (or gazette) of the entity (e.g., list of all college names) starting from a small seed list of example entities.

7.2 Information Retrieval

Information Retrieval (IR) technologies – in the form of search engines like Google™ and Lucene™ – can be used to effectively search resume repositories to locate and retrieve resumes that meet specific criteria. For example, a search engine can retrieve all resumes that contain either of the phrases real-time system or embedded system.

While such searches over resume repositories, though simple to specify and quite efficiently performed, nevertheless have several limitations. For instance, the search engine does not know about domain concepts and relationships like “C++ and Java are both Object Oriented programming languages” and hence a search for Java programmers will not generally return resumes of C++ programmers. A search engine usually has only a limited understanding of natural languages; e.g., it may not know that Programmer and Coder are really the same things. Some searches over resumes require looking for entities which are somewhat semantic; e.g., retrieving resumes of people who have performed the role of automotive hardware tester requires understanding the meaning of role. Finally, some typical searches are simply not possible using search engines; e.g., a search like “Java programmers with at least 5 years of programming experience out of which at least 3 years are in a finance organization”. This search requires computing the total Java programming experience of a person, which may not be explicitly stated anywhere in the resume but can be inferred by analyzing his projects.

The result is that the end-users have to be adept in designing the search conditions, manually analyzing resumes retrieved by the search engine and iteratively refining the search conditions until the required task is done. It is not difficult to see that the end-users would rather prefer a one-shot search mechanism that understands the nuances of their resume search requirements.

7.3 Data and Text Mining for Improving Efficiency of HR Tasks
Data and text mining techniques can be used on raw resume text as well as on structured data extracted from resumes using IE techniques to create summaries, identify trends/patterns etc. that are useful for various purposes; for example:

- **Create technology profiles** (trends/patterns) that can help in optimal skill upgrade planning
- **Create job market intelligence** (trends/patterns that characterize available talent pool) that can help in optimal recruitment planning

Fig. 4 shows some fictitious charts that can be used to obtain useful insights.

Many advanced data mining techniques (predictive modeling, dependency analysis, clustering, anomaly detection, subgroup discovery, root-cause analysis etc.) can be used to derive novel and actionable insights over information extracted from resumes. Many of the use-case scenarios listed in section 4 actually require significant efforts to design domain-driven analytics functions aligned to specific HR business goals.

**8. The ResumeCenter Solution**

TCS is currently building an integrated single-window system, called ResumeCenter, for processing and analysis of resume repositories (both internal and external resumes). ResumeCenter will be integrated in TCS HR Management processes and systems and will provide most of the functionality discussed in this Note as to various HR executives at appropriate points in the HR workflows. ResumeCenter will also make all this functionality available as Web Services, so that other HR systems can programmatically request and access results of resume analysis and processing functionality. The ResumeCenter solution can also (in principle) be made available to external clients outside TCS; e.g., as part of the platform-based solution from TCS for large enterprise customers who outsource complete HR back-office functions. Fig. 5 shows a high-level architecture of the ResumeCenter solution. See [3] for some technical details on the information extraction module in ResumeCenter.
A brief description of the modules in ResumeCenter is as follows.

<table>
<thead>
<tr>
<th>Module</th>
<th>Function</th>
<th>Consumer application/function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resume Extractor</td>
<td>Extracts resume data</td>
<td>Recruitment Portal, Skill repository</td>
</tr>
<tr>
<td>Resume search</td>
<td>Searches the right resume to match job requirements</td>
<td>Recruitment Portal, Skill repository</td>
</tr>
<tr>
<td>Talent Profiling</td>
<td>Trends and profiles of talent available in various organizational entities and external talent pool; Help in implementing organizational strategic talent needs;</td>
<td>Management, Talent Acquisition Planning, Skill repository</td>
</tr>
<tr>
<td>Talent Acquisition Operations Analytics</td>
<td>Offers analysis and monitoring data for measuring efficiency, effectiveness (cost, effort) for the achievement of recruitment objectives</td>
<td>HR Management, Talent Acquisition Management</td>
</tr>
<tr>
<td>Training planning</td>
<td>Offers planning for induction, gap coverage in skills as identified in recruitment of experienced professionals, implementing re-training programmes and for implementing strategic talent achievements in organization</td>
<td>Management, HR Management, Skill repository</td>
</tr>
</tbody>
</table>

9. Conclusions

Repositories of resumes, of both external candidates and current employees, contain valuable information that can be used to improve the quality of work as well as satisfaction levels for employees. It can also be used to improve HR processes (costs, efforts, quality). We identified several use-cases where information and insights obtained from resumes can improve the quality of results. We briefly reviewed several technologies such as natural language processing, text-mining, data-mining and
information retrieval, which are mature enough and can be applied to processing and analysis of resume repositories. We suggested that it is possible to build right set of resume processing and analysis tools, respecting the candidates’ privacy and complying to the applicable data privacy laws, which are closely aligned to each of the identified use-case scenarios and deliver high-quality information and insights required there. One way forward would be to build a central system, which may be called ResumeCenter, which delivers all services related to processing and analysis of resumes to various users and systems across an organization.

References