Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 7, Issue, 04, pp.14427-14429, April, 2015

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

ISSN: 0975-833X

REVIEW ARTICLE

EXTRACTION OF USEABLE INFORMATION FROM UNSTRUCTURED RESUME BASED ON KEYWORDS TO MATCH RESUMES AND JOB PROFILES

***Swaroop Chandre, Swapnil Kad and Viraj Kale**

Department of Computer, Singhad Institute of Technology, Lonavala, India

ARTICLE INFO

Article History:
Received 06th January, 2015
Received in revised form
06th February, 2015
Accepted 20th March, 2015
Published online 28th April, 2015

Key words:

E-recruitment,
CV, analysis,
Information extraction,
HR-XML

ABSTRACT

In the current internet era use of internet in different fields restricts no bounds. Companies and organisation are moving from the traditional recruitment process to E-recruitment. Job seekers or applicants submit their Curriculum Vitae (CV) directly send them to the company's website. At such time company face a lot of problem with these growing number of documents which are in free and different formats. Our work is basically on CV analysis. This paper describes a system for automated resume information extraction to support rapid resume search and management. The describes system is capable of extracting several informative fields describes by HR-XML from free format resume. Experimental results carried on large number of resumes show that the proposed system can handle a precision of 91% and a recall of 88%. The proposed system will be kept in Semantic Web approach that provides companies to find expert finding in an efficient way.

Copyright © 2015 Swaroop Chandre et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

INTRODUCTION

Finding a job on the Internet is quite popular since it enables job seekers to see all the job vacancies and catch career opportunities easily. In addition, companies/institutions are able to hire efficiently qualified employees with the help of online recruitment websites. Storage of millions of resumes in free-structured format in relational databases of the companies is highly time consuming and requires a great deal of human effort.Companies and enterprises receive several hundreds of resumes from job applicants every day.

Literature Survey

A. Existing System

In general there is no standard format in which a resume can be written. To make a standard and search easy enterprises force job seekers to fill an online template. This online template consist different fields to fill in as candidates general, academic and experience. A problem with this approach that the applicant is forced to tune their resume to match the style of the template which might not be able to capture all the details that the applicant might wish to display on their resume. This thus results in tuning the applicants resume as per different websites which also is a time consuming process. Such online template help in building electronic resume database which further helps in fast searching and processing of resumes. Automatic extraction of information from resumes with high precision and recall is not an easy task is there is no specific standard for resume construction. Resumes can be written in many different formats (e.g. structured tables or plain texts) and in different file types (e.g. .txt, .pdf, .doc(x) etc.).

****Corresponding author: Swaroop Chandre,***
Computer Dept, Singhad Institute of Technology, Lonavala, India.

B. Conflict

Another conflict in online resume builder is that the applicant has to map once skills according to the template provided by the website which may not be the applicants priority. Another issue that is taken into consideration is the cost.

C. Substitution

In this paper we describe a system, which is capable of processing resumes in different formats and forms and building an electronic database from the resume. Our system aims in removing the manual effort in screening resumes received by companies to ascertain the suitability of candidate. We organize the system in 2 different sections. We describe the system that is capable of automatically extracting relevant information from a resume and pushing the information into a database. The extraction of relevant information is based on a set of natural language processing and pattern matching techniques. The complete system is web enabled to make it reachable to a large number of people within the company. We discuss the performance of the system in terms of precision and accuracy of the system.

Proposed System

Our proposed system is capable of automatically extracting information from resumes in English language which further populates the structured database. We have designed a system using different modules. The system provides an interface which help in finding suitable candidate of the companies need .The information extraction module is the most significant component of the system. . The information extraction module is capable of extracting important relevant information from a free format resume automatically. The database build module populates the database with the extracted information and builds a resume database. The search field in the system is designed for the organization authority to query the system with his/her need. . However a natural language interface to search resumes to enable searches like "Show me all the resumes that have more than 3 years of java experience". The input module is a web interface which allows input of resume to the system. Our input module is capable of accepting multiple resumes in different forms as of a .zip, .tgz, .7z, .Z, .gz file.

Related Work

The proposed information extraction module is capable of automatically extracting informative fields such as: total experience, date-of-birth, email-id, skill set and qualification from any given resume. This information extraction module uses a natural language processing (NLP) algorithms to extract relevant information from a free format English language resume. The search module provides an interface to query the system for a specific form of resume. The user can query the resume database based on a combination of above mentioned informative fields . All the resumes matching the criteria are displayed with a percentage of criteria matching .The system further provides hyperlink of the location to view complete resume in its original form.

Information Extraction Using NL Techniques

Information Retrieval (IR) systems uses a simpler data model than database systems. Information organized as a collection of documents. Information Extraction locates relevant documents on the basis of user input such as 'keyword'. Keyword based information extraction can be used not only for retrieving textual data, but also for retrieving other types of data such as audio or video associated with keywords. All resumes are unique amongst themselves some may be structured some unstructured hence they are dissimilar.

We one can assume a typical resume to be 2 layered structure. The first layer is composed of several general information blocks such as personal information, education etc. The second layer of structure is within the first layer and contains specific information corresponding to the layer 1. For example, the layer 1 personal information block consists of layer 2 information like name, address and e-mail. Additionally, the location of the information (like name, age etc) in resumes vary significantly from resume to resume. Our system can work on both layered structure and unstructured resumes. The extraction process uses a set of language processing techniques which are part heuristics and part pattern matching.

Modules

Qualification Extraction Module:

Our working system is capable of extracting qualification related keywords from reference data files. These reference data files are created offline. Knowledge-base is capable of extracting qualification keywords in morphological forms like as of Bachelor of Engineering and BE refers to the same thing. Candidates qualification, name of the university and degree class is extracted. If there are multiple qualifications of the candidate they too are extracted.

Skills Extraction Module:

Software skill extraction module helps in searching for skill sets listed in the resume. The system extracts different skill sets like synonyms and morphological skills. Like in all other modules, the system uses a spell checker to identify and resolve typographically errors in the resume. The procedure adopted by the skill extraction module is to initially form n-grams($n < 7$) from the resume.

Experience Extraction Module:

Experience of a candidate is one of the most important fields that is used for identifying candidate for the job. The experience extraction module tries to identify if the candidate has mentioned his total experience like in "I have a total experience of 5 years in this specific industry" by looking for patterns like "total years of experience", "experience total of .... years", "entry level experience", "years of professional experience", "years of experience", "professional exposure of.. .. years/months" etc .. If there is no mention of total number of years of experience in a resume, then the experience extraction module identifies all the time periods spent by the candidate in different projects and then sums up the experience in each project to come out with a total number of years of experience.

Name Extraction Module:

Name extraction module collects all possible names of the candidates and determines the word with highest probability as the candidate name. The processing involves identification of all parts of the resume having the pattern "name" but no "project" or "father" patterns preceding the pattern "name" (to avoid project name and fathers name which are common in a resume). In isolated regions symbols like':', ':-', '-', ',', '_' are used as the delimiter.

Email Extraction Module:

Any line having the patterns like "@", "[at]" become candidate for email-id of the candidate. This mechanism however determines all the email addresses in the resume. A post processing is done by analyzing all the email addresses returned by the system to pick the probable email address. For example, all email addresses collected from the "References" part of the resume are not considered as the possible email id of the candidate. Email id having the candidates name embedded in it can be checked to get best possible candidate email.

EXPERIMENTAL RESULTS

Performance of the system depends on the accuracy and span of the knowledge base. This is applicable to all the natural processing language based system. A total of 100 resumes, mostly of the candidates who were applying to a company, were obtained. A set of 50 resumes, picked randomly, were used to populate the knowledge base. These 50 resumes formed the set of reference resumes. The performance of the system was tested on the remaining 50 resumes. The criteria used for evaluating the performance of the system are precision (p) and recall (r). These metrics are used to measure the performance of any information extraction (IE) systems. Precision and recall metrics are defined as follows:

$$p = \frac{\text{Number of correctly extracted fields by system}}{\text{Total number of fields extracted by system}}$$

$$r = \frac{\text{Number of correctly extracted fields by system}}{\text{Actual number of fields}}$$

Sample output of the Proposed System:

Name: Kishor Kumar
Software Skills: Programming in C,C++, Phyton
Qualification: BE (Karnataka University)- 60%, HSC (State board)-66%, SSC (State board)-81%
Experience: 3+ years
Email: kumarkishor@gmail.com

Future Work

Our further work emphasis on building an approach to extract special skills to improve performance of resume selection. Currently available services filter out thousands of resumes to some hundreds. Since these hundred resumes may be of similar form HR representatives have to search in each resume to find out the special skills in these resumes. Our future work will emphasise the same of finding special feature from some filtered resumes.

Conclusion

In this fast moving world the use of internet bounds no limit. . Candidates post their resume on company's website for recruitment . Further HR representatives have to manually sort the resumes in order to find the right candidate for the job position. This results in an overhead work hence there is need of system which an ease the sorting information from a free format resume for faster recruitment process.

Our proposed paper does the same it aids in searching relevant information from structured as well as unstructured resumes. Proposed system is capable of extracting 6 major fields described by HR-XML. The 6 fields are as extraction of Name, Qualification, Skill, Experience, Email and Date of Birth. Hence this system will reduce the work overhead an and help in the making the recruitment process semi automatic.

REFERENCES

Automatic Extraction of Usable Information from Unstructured to Aid Search, Sunil Kumar Kopparapu, TCS Innovation Labs - Mumbai, Tata Consultancy Services, Thane (West), Maharashtra 400610, Email: SunilKumar.Kopparapu@TCS.Com

Finn, A. and Kushmerick, N. "Multi-level boundary classification for infonnation extraction," in Proceedings of 15th European Conferenceon Machine Learning, Pisa, Italy, September 20-24, 2004

Mining Special Features to Improve the Performance of Product Selection in E-commerce Environment and Resume Extraction System, Sumit Maheshwari 200402041 sumitm@research.iiit.ac.in, sumitmaheshwari.com@gmail.com

Monster, http://www.monster.com

Nahm, U. Y. and Mooney, R. J. "Text mining with infonnation extraction," in Proceedings of the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases, pp. 60-67, Stanford, CA, March 2002.

Naukari, http://www.naukari.com

Towards an Information Extraction System based on Ontology to Match Résumés and Jobs, 1Duygu Çelik, 2Aşkın Karakaş, 3Gülşen Bal, 4Cem Gültunca, 5Atilla Elçi, 6Başak Buluz, 7Murat Can Alevli, Kariyer.net, 2askin@kariyer.net, 3gulsenb@kariyer.net, info@cemgultunca.com.tr,Istanbul Aydin University, 1duygucelik_sw@hotmail.com, basakbuluz@gmail.com, 7canalevli@hotmail.com

Web-based recruiting Framework for CV structuring. Soumaya Amdouni University of Tunis Institut Superieur de Gestion Cite Bouchoucha, Le Bardo, TUNISIA, Soumaya.miagiste@gmail.com Wahiba Ben abdessalem Karaa University of Tunis,Institute wahiba.Abdessalem@isg.mu.tn

****