OSJI OPEN SCIENCE JOURNAL

RESEARCH ARTICLE

Resume Searching to Decide Best Candidate Based on RELIEF Method

Sara Nasr[1]*, Oleg German[1]

[1]Belarusian State University of Informatics and Radioelectronics, Belarus

*Corresponding author: Sara Nasr: sara.nasrh@gmail.com

OPEN ACCESS

Citation: Nasr S., German O. (2020) Resume Searching to Decide Best Candidate Based on RELIEF Method. Open Science Journal 5(2)

Received: 18th January 2020

Accepted: 6th April 2020

Published: 10th June 2020

Funding: The author(s) received no specific funding for this work

Competing Interests: The author have declared that no competing interests exists.

Abstract:

Nowadays the number of people searching for work is increasing, and the number of university graduates is also getting more significant. Companies seeking a perfect job candidate for a person working in the informatics domain can be lost between different resumes with different formats and large quantities. Selecting the best curriculum vitae (CV) is considered a big step to help any company to decide whether a job seeker is suitable or not. The criteria that should be followed to check a CV is to divide it between a block of private information and another block of professional data. The blocks are based on age, education, participating in real projects, availability of published papers or research activities, knowing modern programming languages and technologies, and so on. Selecting the perfect candidate needs an estimation based on a choice function similar to the utility function, which weighs the different criteria and helps evaluate them numerically. But in some cases, the information may be fuzzy or uncertain information, so for this reason, the use of the modified Boyer Moore algorithm and RELIEF method is proposed in this paper to reach the target.

Keywords: Resume, Job description search, Text with mistakes, Short text processing, Searching methods, Fuzzy text, Semantic blocks

Introduction

In general, when searching for some information in a text, the goal is to reach some specific business resolution. Searching CV's for a particular data is one of the problems that online recruiters and many big companies are facing. It is not an easy job for such companies to mine through data of different types and

formats and between a large number of resumes. Previous studies had suggested a system to search for specific information even if it is not precisely stated in the resume. Thus, this study suggested using the modified Boyer Moore algorithm to get targeted information (1). The suggestion is based on the existence of some mistakes in the resume when writing incorrect words without noticing the errors. Another reason for this suggestion is to combine different variations of the same word since they have the same meaning and explaining the same idea.

In the following sections, the issues of fuzzy text and an explanation of how the modified Boyer Moore algorithm can be expressed will be discussed. Section 2 of the paper presents a brief description of the most common approaches for searching short texts and existing applicant tracking systems. Section 3 discusses the proposed method of extending the modern Resume Processing Systems. Section 4 explains the flow chart of the used algorithm and how the text shall be divided into blocks. It also describes the criteria on which the best resume should be selected. In the end, an overall conclusion shall be summed up in Section 5.

Related Approaches

Usually, companies who need assistance in recruitment and in the hiring process use applicant tracking systems (ATS) to assist them. Of course, each of these systems may offer different features and may combine a group of elements that make each system characterized by a particular structure than other systems such as Taleo(2), iCIMS(3), Greenhouse(4). However, all ATS is primarily used to collect, organize, and filter the resumes or job applications. Usually, any candidate applying for a job has to submit his/her CV online and should use such systems as required from companies. Moreover, some recruiting and hiring companies nowadays refuse to receive paper resumes, or email inbox resumes to minimize the mess and to be more organized and efficient. In addition to the above systems, some companies may develop their own applicant tracking systems, especially huge companies such as Google, Microsoft, Apple, and Facebook (5). But even though the resume processing systems mentioned above are successful and proved to be widely used, it is essential to say that these methods don't work with resumes that don't have a predefined structure. Therefore, the use of these systems becomes limited only to companies that ask for a specific predefined structure of resume writing. As a result, this process obliges the applicant to follow many limited rules to submit a resume to one of these companies.

Most computer applications involved in word processing, text searching techniques, or image pattern matching are concerned with pattern recognition. The simplest form of text pattern searching is to compare and match consecutive positions of the text to the pattern, so this way, the comparison will go through all over the text by default from left to right. Of course, it is a successful method when a small amount of text comparison is involved and is called Brute-Force string matching (6). When searching a large number of resumes, Brute-Force technique will be logically inefficient since the amount of data are considerable. Since 1977 when the Boyer-Moore algorithm (6) was created, many approaches and a lot of papers and researches were developed to achieve two crucial criteria - needed for any searching method and pattern matching. Those criteria are the reduction of pattern comparison and, at the same time, the minimization of the processing time needed to do so.

Considering that the traditional applicant tracking systems don't work correctly with distorted texts when the resume includes mistakes or incorrectly stated words, then the resume of the candidate will not be processed successfully. This may happen because these systems use only specific keywords that must be detected in the resume, and in some cases, these words may be miswritten in some candidate's resume or may not be mentioned at all in some other resume, even if the candidate is suitable for the job.

Materials and methods

Based on the above approaches, our goal is to extend possibilities of the modern Resume Processing Systems and make them more intelligible. In this aspect, our approach will search a text, but instead of searching keywords, it provides means to divide the text into a semantic block and recognize the corresponding query answering system to such a block organized text, allowing mistakes in the original text based on Modified Boyer Moore method. So, instead of strictly organized resume representation, the suggested method provides flexibility in resume writing and does not use keywords, which can increase the relevance of the answers to queries prepared in advance. The proposed method also provides means to estimate applicants based on the multi-criteria decision-making process through the implementation of the RELIEF method (7), which saves time for the experts realizing the final selection of the candidates.

Our approach, afterward, is based on three main issues:

- Dividing text into semantic blocks
- Realization of the sequential search in the document with possible mistakes based on Modified Boyer Moore method
- Estimating an applicant accordingly to the multi-criteria decision-making technique using the RELIEF method.

The advantage of the suggested method over already existing fuzzy searching methods for short text is that it can process a text with a higher number of mistakes, and at the same time, it can still find the correct query answer inadequate processing time.

The searching procedure is explained below and can be represented using the flow chart diagram found in figure I:

- When the searching procedure starts, the query and the resume will be processed.
- In each iteration, a CV substring will be compared to the input query.
- If the number of mistakes between 0 and k1 similarity is established based on the Modified Boyer Moore Method. Where k1 is a constant used as a reference or as the maximum number of mistakes that can be found in a substring when compared to a specific pattern(8),(9).
- If the query is found then the comparison will be shifted to the right to search for another existence of the query in the CV.
- If the number of mistakes is higher than k1 we check if it is between k1 and k2. Where k2 is an additional constant added by

the suggested method to permit a more significant number of mistakes in the text.

- If No (the number of mistakes is more significant than k2), then we shift the query to the right to compare to the next substring
- If yes (the number of mistakes is between k1 and k2), then Dice metrics (10) are calculated to compare the query search to the substring. Dice metrics for string comparison since it is tested with heterogeneous data sets, giving less weight for outlier data (11).
- If Dice metrics answer less than 0.5 then we also shift the query to the right
- If the Dice metrics is more significant than 0.5 then the similarity is established
- In all cases, we will shift the query to the right.
- If the CV didn't end, then we search for another existence of the query in the CV and repeat the same iterations otherwise the process will end

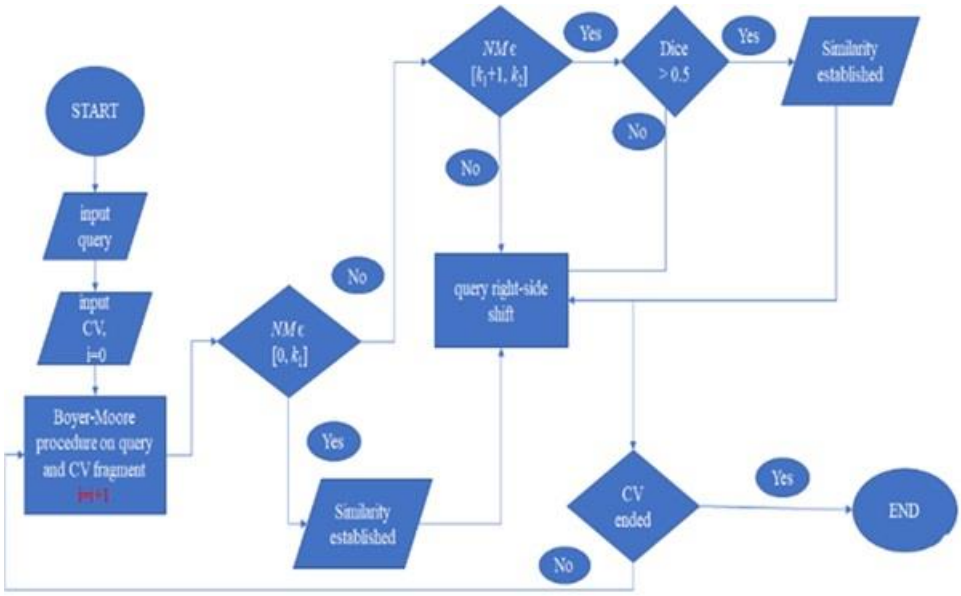

Figure 1: Modified Boyer Moore Search Flow Chart

After implementing the modified Boyer-Moore method, which can detect if a word is found exactly or with some mistakes in a resume, a method is developed so that the resume can be divided into blocks. The importance of blocks division is that our target is to find a complete, understandable statement that explains the needed information that was queried.

To finish the complete process, it should pass through four divided stages (12). At first, any given CV should be converted from its original format, such as a PDF or Word document into a plain text. This procedure can be done by using an already existing system such as Tika. In this way, any additional or unnecessary information (such as colors, fonts) will be removed, and only the raw text will be stored. The second step is after having raw data, get the keyword from the text, and divide them into categories. Since the CV is considered a short text type, it is possible to check all statements of its content and group the

related ideas together based on the keywords similarities. In other words, keywords are selected in a way that similar words are considered as one keyword. For example, based on the CV in Figure 2, the words telecommunication and telecom are considered the same keyword, and each one of them is related to one or more sentences. Noting that a group of pronouns, prepositions, conjunctions, auxiliary and modal verbs (such as can, have, may etc.) are already excluded. In order to compare the similarity of the words, Dice metrics is measured using the formula:

$$P = 2 * (|X| \cap |Y|)/(|X| + |Y|)$$

Doctor Engineer in Signal processing, Electro and Telecommunications

• EDUCATION:

➢ 2006 - 2010: Telecom Bretagne, Brest, France
PhD degree in Signal processing and telecommunications
Entitled: Contribution to the definition of new waveforms for vehicular radars
Keywords: Radar, Signal processing, estimation, detection, performance bounds, statistics and Matlab
➢ 2005 - 2006 Telecom Bretagne, Brest, France
Master, 'Sciences and Technologies of Telecommunications'
➢ 2001 - 2006 Lebanese University, Faculty of Engineering III , Lebanon
Electrical and Electronics Engineering (Telecom and Computer)
➢ 2000 - 2001 : Official college of Beit Jbeil
Lebanese Scientific Baccalaureate (very good)

• PROFESSIONAL EXPERIENCE:

➢ October 2011 – Now : AUL University (Beyrouth and Jadra), Lebanon
Lecturer and Supervisor of 21 CCE senior projects (41 students)
➢ December 2010 – August 2011 : Ecole national d'ingénieurs de Brest (Enib), Brest, France.
Post-doc
Objective: Clustering and Image Segmentation.
Keywords: Image Processing, segmentation, smoothing, Matlab.
➢ December 2010 – June 2011 : Ecole National d'Ingénieurs de Brest (Enib), Brest, France.
Lab assistant: digital signal processing and statistics (CCE 3rd year)
➢ October 2009 – September 2010 : Telecom Bretagne, Microwave department, Brest, France
R&D engineer
Keywords: Signal processing, antenna, Matlab, toolbox building

➢ February 2006 - July 2006 : Telecom Bretagne, Electronic department, Brest, France
Training
Subject: Implementation of fountain codes using c++
➢ July 2005 - September 2005 : Microsoft , Beyrouth, Lebanon
Training
Subject: development of web application using c#.net
➢ 2002 - 2005 : Lebanese University, Faculty of Engineering, Branch 3, Lebanon
• Secured Control with Emergency Procedures, Remotely Accessible Wheel chair
• C and C++ projects : Sets Theory, Numerical Analysis, Data Structures and others
• Java : Simulation of server web pool

• TEACHING ACTIVITIES:

➢ AUL University, Beyrouth , Jadra, and Kaslik, Lebanon (October 2011 – Now)
○ Fiber optics Communications: 108 Course hours (CCE: Masters)
○ Microwave Communications: 72 Course hours (CCE: Masters)
○ Information Theory and Coding: 39 Course hours (CCE: Masters)
○ Digital Signal Processing: 234 Course hours (CCE: BS)
○ Signals and systems: 39 Course hours (CCE: BS)
○ Wireless Communications: 39 Course hours (CCE: BS)
○ Analog communications: 39 Course hours (CCE: BS)
○ Digital Com: 39 Course hours (CCE: BS)
➢ Ecole nationale d'ingénieurs de Brest, France (December 2010 – June 2011) :
○ Signal processing: 21 Tutorial hours
○ Statistics: 10 Tutorial hours
➢ Telecom Bretagne, Brest, France (may 2009):
○ Numerical Analysis: 4 Tutorial hours

• LANGUAGES:

➢ Arabic : Mother Language French : Fluent English : Very Good

• PUBLICATIONS:

➢ International journal:
• Abe, "Estimation Techniques and Simulation Platforms for 77 GHz FMCW ACC Radars", *The European Physical Journal - Applied Physics, November 2011, n° 11001, pp: 1-16.*
➢ International conferences:
• abe, "Strategies for FMCW radars," Proceedings of the 2009 International Transport System technology Symposium, Lille 20-22 oct. 2009.
• Abe, "77 GHz ACC Radar Simulation Platform," Proceedings of the 2009 International Transport System technology Symposium, Lille 20-22 oct. 2009.
➢ French National conferences:
• Abe, "Nouvelle Technique d'identification des caractéristiques spatio-temporelles d'un canal de propagation", JNM 2011, Brest 18-20 Mai 2011.

• OTHER INTERESTS:

➢ Electronics, chess, photography and Photoshop

Figure 2: Resume example

According to the formula, P=2*(|X| ∩ |Y|)/(|X|+|Y|), if two words have the Dice measure equal to 0.5 or greater than 0.5 then the words are considered similar. Taking into consideration a large number of CVs that will be sequentially searched on servers, and noting that they are fuzzy texts, each CV will be processed to produce a collection of two dictionaries. The first combination is a <key, value> pairs, where the key is the keyword, and the value is all the numbers of sentences related to the keyword. The second combination is the <value, text> where the value stands for the number of the sentence, which is related to the key in the first pair, and the text is its related sentence.

To extract the answer of a query, some words of the query is considered keywords. When these keywords are detected, they will be compared to the key part of the <key, value> pair. If there is more than one value combined with the same key, all answers are selected.

Noting that a big semantically block may be selected in some cases may be selected if it is related to the same sentence value and is practically describing the same idea (a whole paragraph).

As a result, we can deduce the following sequential process: keyword → sentence(s) → text block(s)(12).

The remaining part is how to decide which candidate has the best CV. Accumulated CVs must be compared to each other to evaluate the best CV. A choice function $C = \sum w_i.uf_i$ must be used to decide. Since RELIEF method estimates and weights w specific feature for some criteria i, it can be used to get for each feature A, a weighting coefficients W[A], where the features are based on different criteria *i*(13).

Results and discussion

While implementing the above plan, many criteria were taken into consideration. The CVs that were under process were all related to developers' job-seeking candidates. The structure of the CV and the document type was changed so that we could deal with a raw text with an unstructured simple short text document. The application was developed using C# and is still under development for precision and CV comparison. The application was developed first to works with txt extension and then was modified to read different extensions (txt, doc, docx) then extract the raw text from them. The application permits us to obtain some needed information for various candidates. As a first step, the CV is read by the application; then, it is divided into semantic blocks. After the blocks are ready, different queries can be inserted to answer different questions. Some examples were implemented on sample CVs. In Figure 3, three (3) different queries are processed for two (2) different CV of two (2) candidates. The questions of the queries were related to the teaching experience (question 1), post graduate studies (question 2), and the candidate address (question 3). As a result, it has been noticed that the application has successfully detected the correct answers.

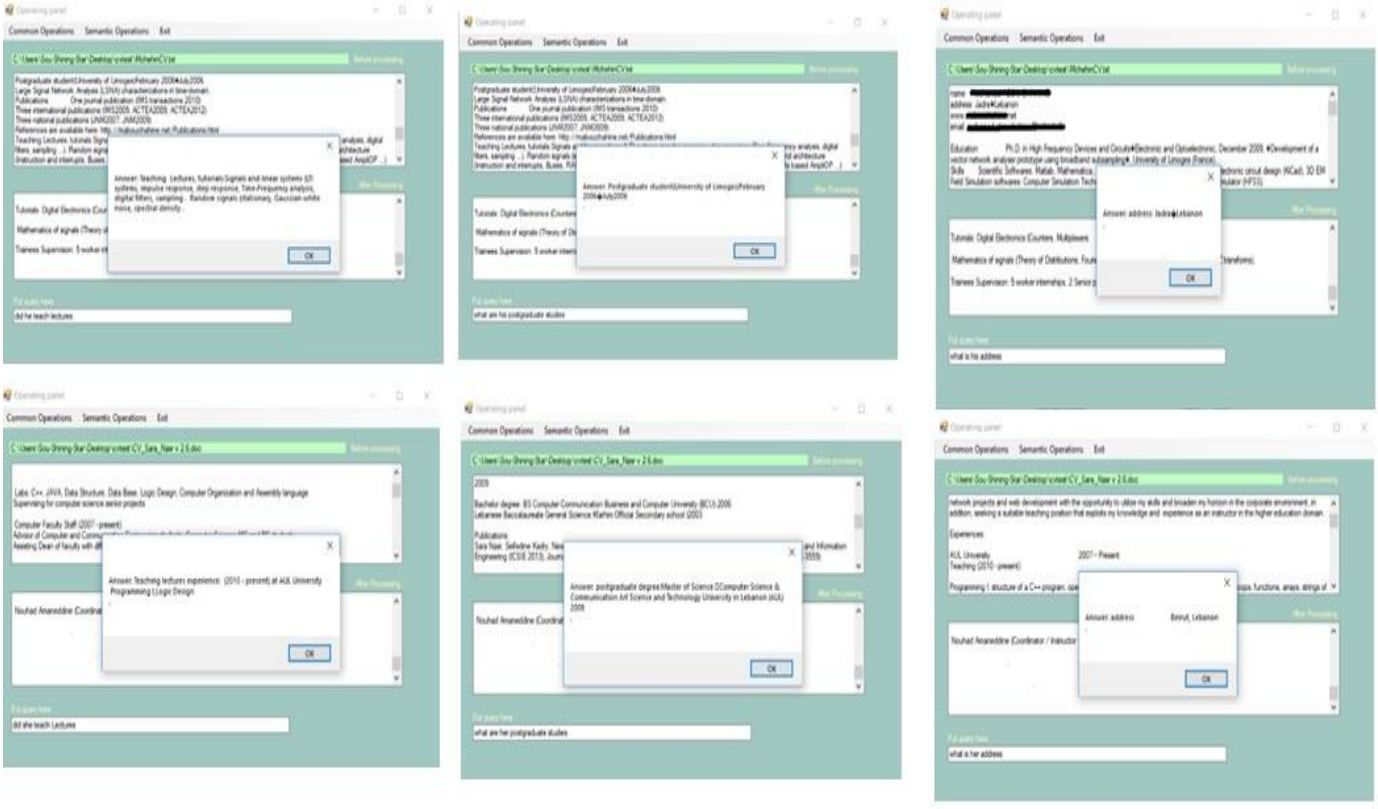

Figure 3: Queries Questions

As an overview of the application the target is to ask different questions concerning the CV in study and estimate the relevance of the answers. For this reason, many CVs were tested by applying them to the application and then dividing them into blocks with this program and ask questions concerning employee interests. Some other sample questions that may be asked are:

1-Have the candidate participated in programming projects?

2-Does he knows modern programming languages?

3-In which university has he studied?

4-Does he has a teaching experience?

Future work and conclusion

A utility function will be used as a future step to complete the whole system's implementation. After getting the answers from different CVs, utility function will evaluate the different selected answers, and numeric criteria will be given to discover the priority of each answer. Through providing a numerical evaluation of each answer, different CVs will be compared and the best candidate will be chosen based on the numeric relation to the collected answers.

Enhanced technological methods are always considered as a need to make CV evaluation easier for companies and recruiters. The suggested method is based on Boyer Moore algorithm to search for a string but with a higher number of mistakes or string differences. The goal is to find an efficient and accurate method for searching for an accurate job description in a system full of resumes,

even if these resumes include some fuzzy mistakes or even if the candidate has used similar words but not the exact keywords found in the query.

The suggested method combines techniques of the modified Boyer-Moore Method with verifying string similarity based on Dice metrics. It provides some important advantages for the suggested method in comparison with the Boyer-Moore method as it gives a possibility to accept several mistakes in the text and, at the same time, preserve the semantics of the original text to find at the end the targeted job description. At the same time, the method can return a complete sentence or paragraph based on the block distribution that is applied to break the CV into different blocks. Some queries will be answered, and the answers will be given in the form of blocks. The utility function should estimate the percent of relevant answers. The main criterion is the relevance of the answers to help us choose the best candidate for the targeted job description. RELIEF theory is usually used with exact text since it is an exact method. Through the suggested method, the RELIEF theory will be modified based on the utility method to discover the best answer and, hence, choose the best candidate.

References:

1.Nasr S, German OV. Assessment of Graduate Students' Resumes Using Short Text Searching Method. In: 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE) [Internet]. Sardinia, Italy: IEEE; 2019. p. 306–8. Available from: https://ieeexplore.ieee.org/document/8791695/

2.Taleo applicant tracking system [Internet]. Available from: https://www.applicanttrackingsystems.net/oracle-taleo/

3. Icim applicant tracking system [Internet]. Available from: https://www.icims.com

4.Greenhouse applicant tracking system [Internet]. Available from: https://www.jobscan.co/blog/greenhouse-ats-what-job-seekers-need-to-know/

5.Misa N. 5 Huge Employers That Use Homegrown Applicant Tracking Systems [Internet]. The Magnet recruiting-software. 2016. Available from: https://blog.ongig.com/recruiting-software/5-huge-employers-that-use-homegrown-applicant-tracking-systems/

6.Jayanta Y, Borah P, Talukdar G. A comparison of String matching algorithms-Boyer-Moore algorithm and Brute-Force algorithm. Indian Journal of Science and Technology. 2013 Mar 1;6:pp176-182.

7.Arauzo-Azofra A, Benıtez JM, Castro JL. A feature set measure based on Relief. :7.

8.Chen J, Zhang C, Niu Z. A Two-Step Resume Information Extraction Algorithm. Mathematical Problems in Engineering. 2018;2018:1–8.

9.Galil Z, Giancarlo R. Data structures and algorithms for approximate string matching. Journal of Complexity. 1988 Mar;4(1):33–72.

10.Mandel I. Cluster analysis. In: Finances and Statistics. 1988. p. 176.

11.McCune B, Grace JB, Urban DL. Analysis of ecological communities. 2nd printing. Gleneden Beach, Or: MjM Software Design; 2002. 300 p.

12.Yulia German, Oleg German, Nasr S. Information Extraction Method from Resume. Proceedings of BSTU. 2019;64–9.

13. German Y, German O, Nasr S. Decision making based on relief algorithm. In: Proceedings of the International Conference of Information Technologies and Systems (ITS) 2019. BSUIR, Misnk, Republic of Belarus: BSUIR; 2019. p. 194–5.