

УДК 001.51:004.891

T. Shatovska, I. Kamenieva

Kharkov National University of Radioelectronics, Kharkov, Ukraine
 tanita_uk@mail.ru, irina.kamenieva@gmail.com

Recruitment and Intelligent System

The Carrier Centre is information, analytical and organizational support of job placements of students and graduates. The information system for supporting all main activities was developed. Nowadays the system strengthens links between students and companies as repository of the CVs and vacancies. On the other side the system should be as a virtual recruiter that take into account student’s personal abilities and preferences, available jobs, Company profiles, local labour market infrastructure, industrial and technological trends, account job specification, available human resource to provide the effective decisions on employment. This paper presents the intelligent management system based on text mining methods for supporting recruitment services.

Introduction

Nowadays one of the most frightening social problems in Ukraine is a formal unemployment among young people. Even after receiving the university degree young professionals quite seldom can find jobs which are adequate to their taken specialties. In particular, it is rather complicated to find an appropriate position after graduating from the universities. Closer co-operation between the universities and the enterprises is needed during strategic planning of the new high-tech positions in companies. Such a co-operation would be beneficial for the both sides and will allow students to get high quality positions in private and public sectors. Two main classes of services provided by Carrier Centre: to help educated professionals to find appropriate job and to help companies to find right professionals to fill available jobs.

The University Carrier Centre provides a student’s consultancy and solution taking into account his/her personal abilities and preferences, available jobs, account job specification, available human resources, pretenders’ profiles based on University diplomas, University and Company profiles, National educational policy and standards, local labour market infrastructure, industrial and technological trends. On the other hand the University Carrier Centre does a lot of labour market research like analysis of students’ placement through practical bases, analysis of specialists’ and masters’ employment to enterprises, weekly statistics of student’s employment, analysis of tendencies of labour demand for specialties. This means that Carrier Centre makes analytical research and proposes solutions for student competitiveness growth (fig. 1). Of course this is not possible without an appropriate modern analytical information system.



Figure 1 – Stages of competitiveness growth

One of the system module is *virtual consultant for employment*.

The CV description is a challenge. Each of us is individual and delivers the information in one's own way. The manager of the University Career Centre looks through and analyzes daily tens resumes of the students, faces the most different styles of their writing. Formatting, fonts and logic structure of the CV's are completely arbitrary. Moreover some companies have their own structure of CV's, but some of them want to review and analyze the style and logic of the students' CV's who do not have big experience in writing them. Systems which allow shifting on itself a part of the most routine actions of the manager on processing of the CV's and companies vacancies which are urgent now. Two main features of such system are: automatic information gathering about candidates and vacancies of the companies, clustering CV's and vacancies, and automatically matching the most favourable vacancies to the corresponding CV. In other words this system must work as a virtual intelligent web-consultant for the student. It is necessary to notice that there are criteria on which CV of the student is selected by companies but it is subjective approach.

Nowadays the process of comparing the new CV's from the earlier existing one, classification of CV's is completely by hand. The number of CV's grows (and the number of CV's copies of the same candidate) and admissible time for their processing is reduced. It leads to that it is impossible to process the whole stream of arriving CV's when their processing is made manually. Creation of the intelligent web-system as a virtual recruitment consultant allows to solve the following problems:

1. To make the analysis of structure and recognition of fields of the CV for formalized representation.
2. Automatic gathering of the CV's and vacancies from certain web-sites and their addition into database.
3. Classification of all CV's and vacancies according to the subjects.
4. Elimination of duplicates.
5. Flexible search according to user's inquiries.
6. Ranking of the resume and vacancies inside group: taking into account existing hierarchy of a subject domain using a matrix of skills and abilities.
7. Matching vacancies to the most corresponding CV's of the students.
8. Annotation of the CV and CV's groups.

1. Overview

Some part of the information about the candidate is now carried out manually that leads to information distortion. However the automatic information extraction is not always correct, therefore completely automatic mode does not approach the given problem. More suitable is the automated mode with manual acknowledgement. In the majority of cases automatic processing yields good results but in case when the system could not correctly distinguish some parts of the resume the manager carries out a marking manually. Thus, the system receives one more copy of training sample which will be used in the next training phase. Also the system should be able to check the consistency of the resume. For example, it is used to check the intersection of job periods in different places, to check skills in CV and in the description of real projects. As a basis function the system should classify the arriving resumes and vacancies to certain groups and update a database.

Unit patterns of the CV are not fixed; therefore generally we consider that the candidate and companies send the resume and vacancies in any form. CV of the candidate always consists of some parts, in other words, has a logical structure. Logic blocks usually

are called. It allows to allocate them in text for text clustering methods [1-4]. In spite of the fact that styles of a CV's writing strongly vary general blocks can be extracted like Surname, Sex, Date of birth, Descriptions of the previous job places etc. Therefore we allocate common set of attributes which could be present in the majority of the CVs.

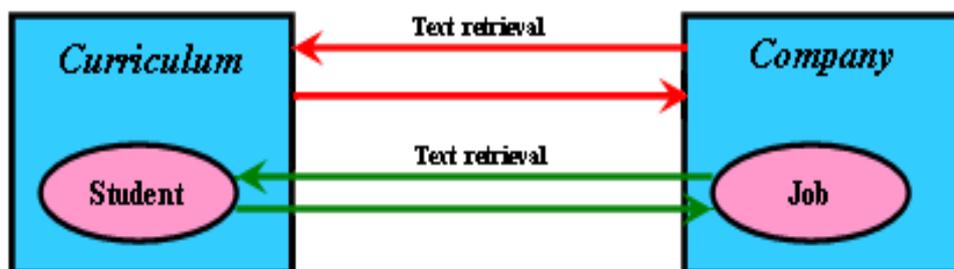


Figure 2 – Text information analyse

To provide the module flexibility the fixed templates and rules of data extraction from CV and vacancy are not used. The model of the CV and vacancy is created. We use the approach to *adjust* each resume to the constructed model of the CV. For some final objects the set of rules for information extraction should be created. If some blocks are incorrectly distinguished the module turns to the training mode and creates the additional rule of the information extraction which is put down to the knowledgebase. When some new CV's blocks appear (for example, the information about additional interests) in a mode of model editing the new elements are added and the extraction rule is constructed. Then all CVs are updating with the link to a new property. This module can be adjusted on extraction of any data from job placement area – the resume, the analysis of questionnaires etc.

2. Method descriptions

2.1. Curriculum Vitae Clustering

Clustering is a process of grouping the data into classes or clusters so that objects within a cluster have high similarity in comparison to one another, but they are very dissimilar to objects in other clusters. The list of classes is defined in advance and includes all necessary directions of student's education of our University: management, mobile communication, computer science, radioelectronics etc. Each resume after processing is presented in the form of such scheme: a key-value, as $R = \{r_i\}$, where r_i – the resume, $r_i = \{< \text{key}, \text{value}>\}$, where $i = 1 \dots n$, n – number of attributes.

The description model is the same for all CVs. For each cluster the rule of the resume frequency in determined group was define as $F = \{f_{i1}, \dots, f_{im}\}$.

Applying the given conditions, we will receive set of intersecting subsets $C_i \cap C_j \neq \emptyset$, where $C_i = f_i(r_i)$. It is shown on Fig. 3.

For each attribute the measure TF-IDF was applied [2], [3]. Each CV or vacancy d is considered to be a weighted vector in the term-space and each document (vacancy or CV) can be presented as $tf_1 * \log(n/df_1), tf_2 * \log(n/df_2), \dots, tf_m * \log(n/df_m)$, where tf_i – is a frequency of i -th term in the document and df_i – is a number of documents which contains i -th term, and n – is a total number of documents in sample. Each vector of the document should be normalised, $\|d_{tfidf}\|_2 = 1$.



Figure 3 – Initial clustering

For similarity estimation the cosine similarity method is used, which is defined as in equation (1).

Similarity measure

$$\text{cosine}(d_i, d_j) = \frac{\langle d_i \times d_j \rangle}{\|d_j\|_2 \times \|d_i\|_2} = \frac{\sum_{t=1}^t d_i \times d_j}{\sqrt{\sum_{t=1}^t d_i^2} \times \sqrt{\sum_{t=1}^t d_j^2}}, \quad (1)$$

where d_i and d_j – components of vector documents, t – is a vector dimension. Further it is calculated the total TF-IDF weight to classify the CV and vacancy.

2.2. Clustering the CV's using integrate approach

Inside each cluster we define the conditions to separate CVs and vacancies into subclusters to group a subcategory. These grouping conditions are defined by the user. In other words each resume or vacancy is an object with attributes where attributes are properties of the resume or vacancy, for example the description of certain skills. At the first stage of clustering we use a hierarchical approach [5], [6]. It creates a hierarchical decomposition of the given dataset.

We integrate hierarchical agglomeration and iterative relocation by first using a hierarchical agglomerative algorithm with UPGMA method [6], [7] and then refining the result using our iterative approach [8], [9], similar to the Chameleon clustering [10]. At the final iteration of algorithm it determines the similarity between each pair of clusters by taking into account both at their relative inter-connectivity and their relative closeness.

In our algorithm during first phase we construct an asymmetric k-NN graph and there exists an edge between two points if for one of it there exists closest neighbour among all existing neighbours according to the value of k. Note that the weight of an edge connecting two objects in the k-NN graph is a similarity measure between them, as usual a simple distance measure (or inversely related to their distance).

The weight of an edge we compute as a weighted distance between objects. During coarsening phase the set of smaller hypergraphs is constructed. In the first stage of coarsening process we choose the set of vertices with maximum degrees and match it with a random neighbour. On the other stages we visit each vertex in a random order and match it with adjacent vertex via heaviest edge. Note that usually the weight of an edge connecting two nodes in a coarsened version of the graph is the number of edges in the original graph that connects the two sets of original nodes collapsed into the two coarse nodes. In our case we compute the weight of the hyperedge as the sum of the weights of all edges that collapse on each other during coarsening step. We stop the coarsening process at each level as soon as the number of multivertices of the resulting coarse hypergraph has been reduced by a constant less than two.

On the next level of algorithm we produce a set of small hypergraphs using k-way multilevel paradigm [11]. We start the process of partitioning by choosing k most heavily multivertices, where $k = 8, 16, 32$. After that we gather one by one all neighbours from each chosen vertex and obtain the initial partitioning w.r.t to the balancing constant. The problem of computing an optimal bisection of a hypergraph is NP-hard. One of the most commonly used objective function is to minimize the hyperedge-cut of the partitioning; i.e., the total number of hyperedges that span multiple partitions [11]. In our experiments we use a greedy refinement algorithm developed by George Karypis [11], but as the gain function for each vertex we compute the differences between the sum of the weights of edges incident on vertex that go to the other partition and the sum of edges weights that stay within the partition. We choose the vertex with maximum positive gain and move it if it results in a positive gain, so we work only with boundary vertices.

After partitioning of hypergraph into the large number of small parts we start to merge the pair of clusters for which both relative inter-connectivity and their relative closeness are high. In our research we use George Karypis formula to compute the similarity between sub-clusters and modified expression by changing the relative inter-connectivity to a new expression that estimates the average weights of edges in each sub-graph and the number of edges that connect two partitions to the number of edges that stay within the smallest partition. Experimental results showed that this method is not sensitive to the value of k and doesn't need a specific k-nearest neighbour graph creating [7].

The resume is distinguished if the subcluster is defined and in appropriate way is saved into system.

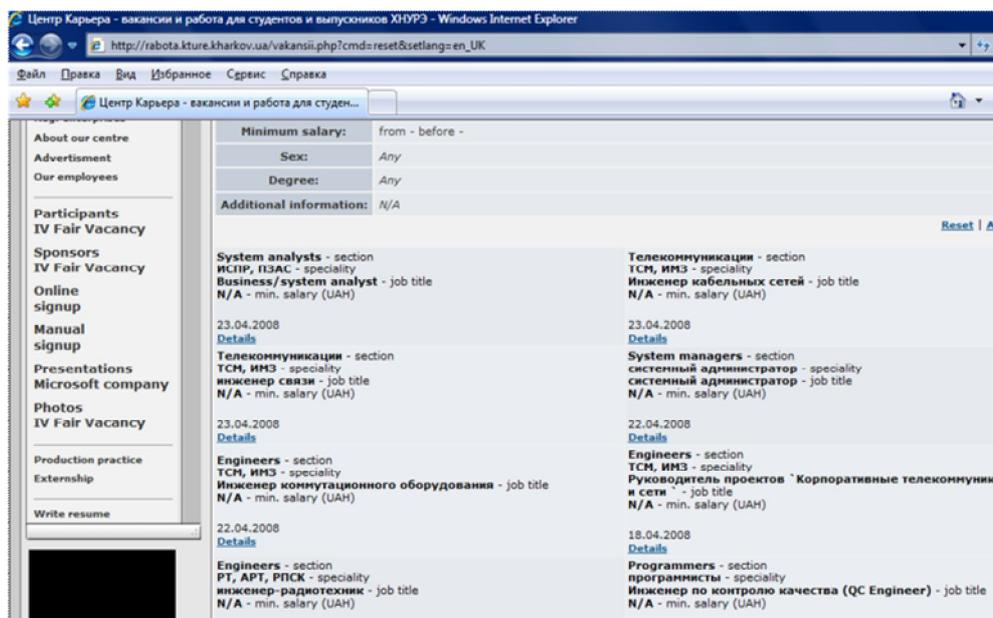


Figure 4 – Clustering of vacancies

2.3 Annotation of the candidate's CV

The systems of texts processing used different approaches to the text annotation. The most widespread way is the list of keywords. This way is simple for implementation, but there is lack of self-descriptiveness. Another way is an automatic abstract construction. This way

gives enough clear abstract, but is combined algorithmically. Considering a problem domain, it is offered to annotate (among other methods) the candidate's CV by adding a subset of the attributes from common blocks of all CVs like skills and abilities of candidates.

In our experiments 200 resumes have been manually selected and marked. The marking included allocation of blocks "Education", "Experience" and "Other". The sections "Contact information", "Hobby" etc has been entered into block "Other". In this part of our experiments we wanted to create a list of keywords for each of described above blocks. At the initial step of the preprocessing not printed symbols, stop words, marking symbols and also superfluous blanks, numbers and abbreviations were deleted. The second step is stemming. The algorithm "Porter stemming" adapted for Russian language is used [12]. It is also simple for implementation as it is constructed on heuristic rules of truncation of words and does not demand dictionary support. Unfortunately, in atypical words it commits errors, but it occurs seldom and does not influence on the final result. After normalization the word is located in the list of keywords for the given block. For the generated keywords their frequency of occurrence is calculated as well. Sometimes it is mistaken and the lines of length less than three appears. As these lines do not concern semantic carriers of the block, it is possible to remove them. As the result of the second step we receive the list of bases keywords with frequency of their occurrence in the block text.

After forming this list it has turned out that one word can belong to several lists. In this case it is not the unique characteristic of the block. For uniqueness maintenance it is necessary to get rid of keywords intersection. Frequencies of such words were compared. The word remains in that group where frequency of its occurrence is more and leaves the group where frequency of it occurrence is less. Thus we have received not intersecting sets of keywords with frequency characteristics for each group.

As practical experiments have shown, using only root of keywords does not give enough exact splitting. Therefore to define the boarder of blocks the phrases of blocks headings were used. At a stage of manual parsing of entrance files headings were separately allocated. In parallel for each of blocks the list of keywords and headings was formed. The accuracy of blocks separating using only keywords of headings was 80 %. If the heading border has not been distinguished the information was saved in system. By heading analysis we have defined possible splitting of the text in CV into blocks. To confirm or correct this splitting the statistical approach on the basis of the available information about keywords is used. As a result we receive the text broken into blocks **Educations, Experience and Other**. Splitting of the text into blocks occurs as follows. For each word we find its normal form using already mentioned algorithm stemming [12]. Further we search for the received normal form in lists of keywords and if we find it we **paint** this word in colour of group.

In Computer Science the fact is the individual value of the data created or used by business process. The facts of our problem domain are: the date of birth, the marital status, date of receipt and the termination of one/several educational institutions, professional skills, citizenship, languages etc. Allocation of the facts occurs by the certain rules, constructed on the basis of regular expressions [13]. They are formed in the training phase.

Using the CV as common model allows reaching the high quality of classification. The offered approach applied to analyze the vacancies and allows to solve such problems as comparison of the CVs and existing vacancies in the system. The method of discovery a list of top resumes (i.e. on what there are a lot demands of employers) is used.

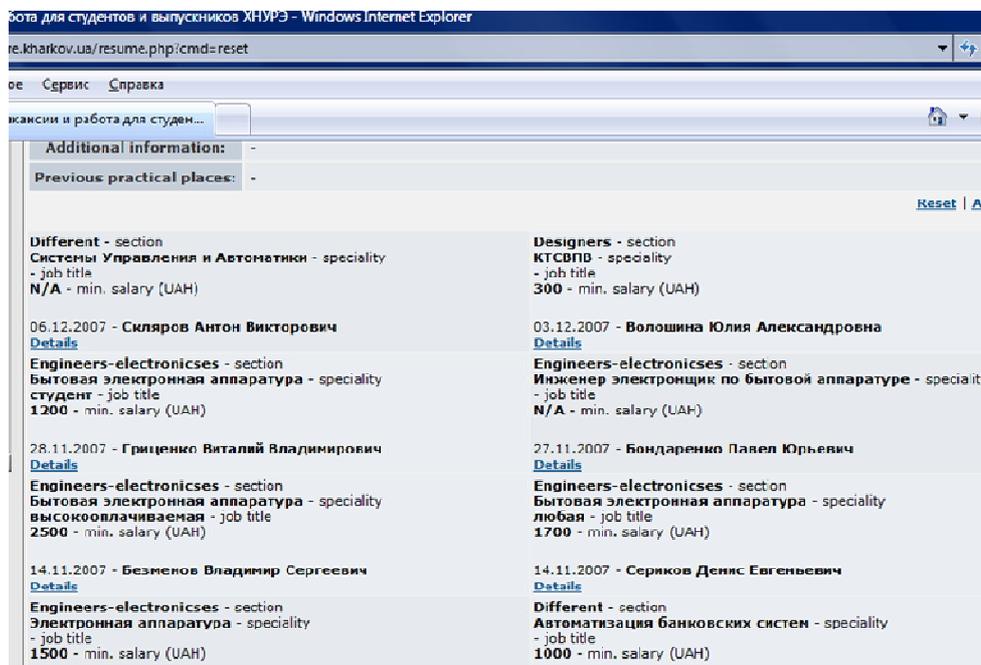


Figure 5 – Top CV's

The pilot version of the system is developed. The efficiency analysis on 500 CVs has shown that in 87 % of cases it has made correct splitting into blocks. And in 82 % of cases the facts have been correctly allocated. The analysis of cases when the system could not break the text into blocks correctly has been carried out: atypical styles of CVs, HTML-tables.

Conclusion

The idea of developing the intelligent system for supporting the employment process of student is offered. As *virtual consultant for employment* it can force the process of job finding.

The offered approach can be used for the decision of classification problems, segmentation and allocation of the facts in other areas connected with document circulation in recruitment services.

Universal model of the CV and vacancy allows to attain high quality of classification. The offered method allows to solve such tasks as annotation of the resume of the candidate and automatic comparison of the resume of existing vacancy. The integrated clustering approach for CV's similarity estimation is offered. On the base of it the list of top actual CV's is formed.

References

1. Liu B., Lee W. S., Yu P., and Li X. 2002. Partially supervised classification of text documents. ICML-02. – Salton, G. and McGill, M. Introduction to Modern Information Retrieval. McGraw-Hill. – 1983.
2. Yang Y. and Pedersen J. P. A comparative study on feature selection in text categorization. ICML-97. – 1997.
3. Andrew McCallum, Rosenfeld R., Mitchell T, Ng A. Improving text classification by shrinkage in a hierarchy of classes // In Proceedings of the International Conference on Machine Learning (ICML) – 1998. – P. 359-367.
4. Toutanova K., Chen F., Popat K., and Hofmann Th. Text classification in a hierarchical mixture model for small training sets. // In Proceedings of the Tenth International ACM Conference on Information and Knowledge Management (CIKM). – 2001.

5. Joachims T. A probabilistic analysis of the rocchio algorithm with TFIDF for text categorization, // In Proc. Of the ICML'97. – 1997. – P. 143-151.
6. Zhao Y. and Karypis G. Evaluation of hierarchical clustering algorithms for document datasets // In Proceedings of the International Conference on Information and Knowledge Management. – 2002.
7. Zhao Y. and Karypis G. Empirical and theoretical comparisons of selected criterion functions for document clustering // Machine Learning. – 2004. – 55(3).
8. Shatovska T., Safonova T., Tarasov I. A Modified Multilevel Approach to the Dynamic Hierarchical Clustering for Complex types of Shapes. Lecture Notes in Informatics (LNI) // Proceeding. – 2007. – Vol. P-107 – P. 176-186.
9. Shatovska T., Safonova T., Tarasov I. The New Software Package for Dynamic Hierarchical Clustering for Circles Types of Shapes // Proceedings of XIII-th International Conference KDS. – 2007. – Varna (Bulgaria). – P. 125-129.
10. Karypis G., Han E.H., Kumar V. CHAMELEON: A hierarchical clustering algorithm using dynamic modeling, IEEE Computer: Special Issue on Data Analysis and Mining. – 1999. – Vol. 32(8). – P. 68-75.
11. Karypis G. and Kumar V. Multilevel k-way hypergraph partitioning // Proceedings of the Design and Automation Conference. – 1999.
12. Russian stemming algorithm, 2005 [Электронный ресурс]. – Режим доступа: <http://snowball.tartarus.org/algorithms/russian/stemmer.html>.
13. Keleberda I., Repka V., Biletskiy Y. Building learner's ontologies to assist personalized search of learning objects. ICEC 2006. – P. 569-573.

Т.Б. Шатовська, І.В. Каменева

Рекрутингова та інтелектуальна система

«Кар'єра – Центр» – це інформаційна, аналітична і організаційна допомога в працевлаштуванні студентів і випускників. Була створена інформаційна система для підтримки всіх основних видів діяльності. В даний час система зміцнює зв'язки між студентами і компаніями як сховище резюме і вакансій. З іншого боку, система повинна бути як віртуальний рекрутер, який бере до уваги особисті здібності і переваги студента, доступні робочі місця, профілі компанії, місцеву інфраструктуру трудового ринку, індустриальні і технологічні тенденції, рахує специфікацію роботи, доступний людський ресурс, щоб забезпечити ефективні рішення у сфері зайнятості. Ця стаття представляє інтелектуальну систему управління, засновану на методах обробки тексту для підтримки рекрутер-сервісів.

Т.Б. Шатовская, И.В. Каменева

Рекрутинговая и интеллектуальная система

«Карьера – Центр» – это информационная, аналитическая и организационная помощь в трудоустройстве студентов и выпускников. Была создана информационная система для поддержки основных видов деятельности. В настоящее время система укрепляет связи между студентами и компаниями как хранилище резюме и вакансий. С другой стороны, система должна быть как виртуальный рекрутер, который принимает во внимание личные способности и предпочтения студента, доступные рабочие места, профили компании, местную инфраструктуру трудового рынка, индустриальные и технологические тенденции, считает спецификацию работы, доступный человеческий ресурс, чтобы обеспечить эффективные решения в области занятости. Эта статья представляет интеллектуальную систему управления, основанную на методах обработки текста для поддержки рекрутер-сервисов.

Статья поступила в редакцию 07.06.2008.