2011 International Conference on Information and Intelligent Computing
IPCSIT vol.18 (2011) © (2011) IACSIT Press, Singapore

Text Mining with Information Extraction in Career Counseling

Rehmat Ullah[1], Khawaja Muhammad Yahya[3] and Samir H. Abdul-Jauwad[2]

[1,3]Department of Computer Systems Engineering, Univ. of Engineering & Technology, Peshawar, Pakistan
[2]Department of Electrical Engineering, King Fahd Univ. of Petroleum & Minerals, Dhahran, Saudi Arabia
{[1]rehmatkttk, [3]yahya.khawaja}@nwfpuet.edu.pk and [2]samara@kfupm.edu.sa

Abstract. As we go "back to university" with this issue of *Computer Science or Engineering Today,* it may be useful to reflect on what we are going to school for [1]. For students, university usually means gainful employment and for their instructors/universities it means, or at least it should mean, getting their students employed. But what do employers want from students? Alumni, meetings with employers and attending industry conferences provide some answers. Students themselves may learn something from job interviews, whether successful or not. But such information is neither structured nor disseminated very well [2]. In this paper we made an effort to present such information in a more structured way using Text Mining techniques. We take the view that employers state what they want in their recruitment advertisements. So, we have analyzed 1000 ads for Computer Science and Engineering jobs. For the purposes of this article, the term *"Engineering"* encompasses Electrical, Mechanical, Civil, Computer, Telecom and Chemical. We interviewed more than 500 Computer Science and Engineering graduates and professionals and also conducted an online survey in which some 1000 people participated from different parts of Pakistan. We did this to make sure that our sample is random enough and not biased towards one group or category. Here we share the results of our analysis to offer insight on job market for Computer Science and Engineering students.

Keywords: Text Mining, QDA Miner©, Job advertisements, Computer Science, Engineering

1. Introduction

This paper addresses the following question ***"What does the Industry wants from Computer Science & Engineering Graduates?"*** The goal of this publication is to provide a way for undergraduate and master's Computer Science and Engineering program directors and instructors to have a detailed understanding of the job market. They can use this understanding to evaluate the extent to which they are providing their students with the necessary training to enter the workforce. Students can use our analysis to supplement their formal education at university and to be prepared for job interviews and jobs. Government agencies may use our analysis to see the production and consumption of graduates and to design policies in order to balance the equation. Finally, recruiting managers can look at what skills, etc. other companies are looking for and use this information to refine their job advertisements. Although all advertisements we analyzed were from Pakistan, a quick check of advertisements from the United States, Middle East, Indian and Malaysian based companies indicate that these companies are looking for the same set of skills and backgrounds. So our results could be applicable to other countries as well.

We have tried to make broad qualitative inferences from our analysis. Afterall, we are analyzing text not numbers, and the text could have a different meaning than what we are inferring. We share these results because we believe that they are useful.

1.1. Our Dataset

Our sample comprises job advertisements from Pakistan-based industry employers. Some Government jobs, in particular from the defense forces and intelligence services, are also included. We collected

advertisements from two of Pakistan's top ranked online hiring portals [Rozee.PK], *Pakistan # 1 Job Website* and [BrightSpyre.COM], *Online Hiring Solution*. We selected placements only that were posted between March15, 2010 and September 20, 2010. At [Rozee.PK] we selected jobs under "Engineering" , "Software & Web Development", "Network Administration & Automation", "Telecommunication & ISP", "Graphics & UI Design", "Production & Manufacturing", "Product & Project Management", "Real Estate & Construction" and "Maintenance & Repair". At [BrightSpyre.COM], we searched for advertisements using the keywords "Engineering", "Computer Science", "Software Development" and "Web Development". This search returned more than 1000 advertisements. The search results were stored in a database. Some employers had posted the same advertisements multiple times so duplicates were deleted manually from the database. There were also quite a few advertisements that were identical except for date of advertisement. Those were also deleted. The end result was 1000 advertisements from both [Rozee.PK] and [BrightSpyre.COM].

Also some 500 Computer Science and Engineering graduates and professionals were interviewed and the result was stored in the database. An online survey was conducted, in which some 1000 people participated from different parts of Pakistan. The result was also stored in the database.

It can be claimed that our sample is typical or random and no category or group is over- represented. It is really a convenient sample. It is hoped that the large number of advertisements, interviews and surveys provided indicative results.

2. Methodology

As in all data mining tasks, the acquisition of the dataset and then pre-processing was the major task in this project [3, 4]. First, the data was acquired and stored in a database. The search was delimited in each instance by a delimiter for the QDAMiner© to store the text for each job in a new case. Likewise to extract numerical and alphanumerical variables from text each variable in the data was delimited by a start and end delimiter. A spell-check was run on the entire dataset to correct any spelling errors and typos.

The text of these advertisements was analyzed, tallying up relevant phrases and keywords in the advertisements. For instance, if the phrase "bachelor's degree" (or variants like B.S. degree) appears in an advertisement, the advertisement is tallied up as a relevant case in the "Bachelor" category. The same advertisement may specify "C++" so it was tallied up in the "Programming" category as well. The approach is to draw out inferences from how many advertisements belong to each category. Each category is comprised of many keywords or phrases, and an advertisement containing any of these keywords or phrases is said to belong to this category.

A QDA Miner© was used, a qualitative data analysis tool from Provalis Research, to list out the frequency of all keywords/codes in all the advertisements. Advertisements were categorized within thirteen master categories: (1) company name, (2) job type, (3) gender required, (4) age range, (5) job posting date, (6) category of job, (7) job location, (8) experience required, (9) salary range, (10) career level, (11) qualification required, (12) discipline background and (13) skills required. There were six job types - contract, permanent, full-time, part-time, freelance and internship. There are five qualification categories – master, bachelor, intermediate/A-level, diploma, and matric/O-level corresponding to the employers' requirements for a position. Likewise there were seven categories for the disciplinary background needed, four different career levels, eleven different ranges of salary, eight major job locations and five different categories of jobs.

The categories within a master category are neither mutually exclusive nor exhaustive. An advertisement can fall in more than one category within a master category. For instance an advertisement seeking someone with either a bachelor's or a master's degree would fall in both the bachelor's and the master's categories even though the lists of keywords corresponding to the two categories are mutually exclusive.

2.1. System Architecture and Data Flow

The system architecture is shown in Figure 1. The data was collected from [Rozee.PK], [BrightSpyre.COM], interviews and survey and necessary pre-processing was performed. The clean dataset was then fed into QDA Miner©. The QDA Miner© has different analysis tools to analyze qualitative data. The *code co-occurrences/case similarity analysis, and the code frequency analysis and variable statistical analysis* were selected to get the desired results.

3. Analysis

3.1. Qualification Requirements

As shown in Figure 2 (a) the majority of high profile companies that placed job advertisements required a bachelor's degree. The analysis signals that one need to have at least bachelor's degree to get into a good job. The number of advertisements in which an intermediate/A-level was sought was only small and insignificant. For some of jobs, a bachelor's degree was required and a master's degree was also sought or was preferable. Similarly for quite a few jobs, a master's degree was required and a bachelor's degree was also acceptable. Recall that these categories overlap in the sense an advertisement asking for a bachelor's or a master's degree would be tallied up under both. Figure 2 (b) and (c) shows this fact using Jaccard's coefficient.

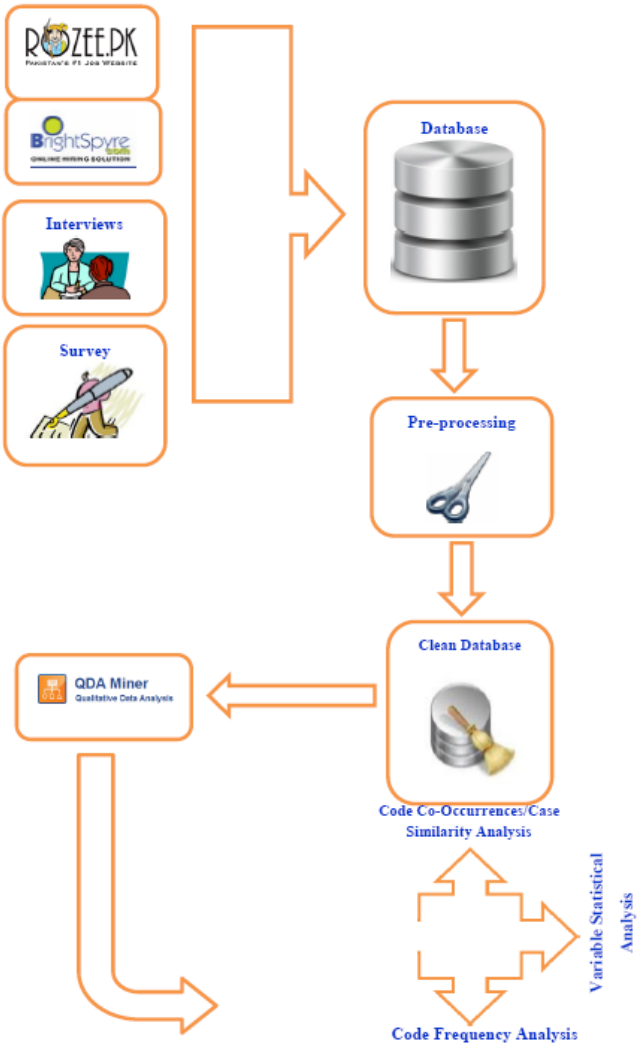
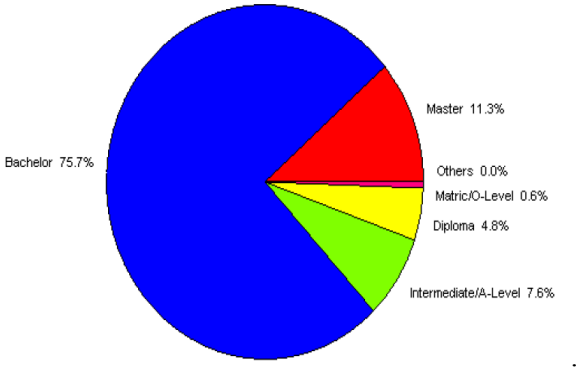
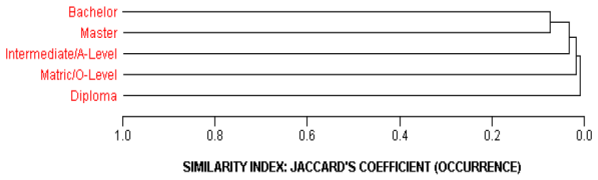


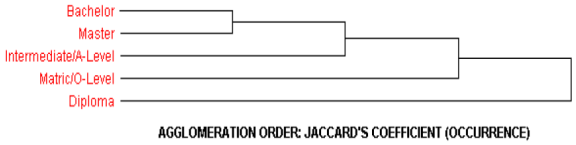
Figure 1 . System architecture and data flow



(a)



(b)



(c)

Figure 2 . Qualification requirements

3.2. Skills Requirements

Companies require a variety of skills and analyzing the advertisements has been extremely useful in building an inventory. Educators, students and employers are advised to pay special attention to these categories. Indeed, the bulk of our analysis effort went to the selecting more than 50 skills related keywords and phrases from which we constructed 10 categories.

Of these categories, some are what might be considered "hard" skills such as "programming" and "database" skills that might already be a part of many Computer Science or Engineering programs. Having such skills would benefit Computer Science or Engineering graduates given that almost a quarter (22.3 percent) of the advertisements requires database skills and 16.3 percent require programming skills. About 5 percent of all advertisements require basic IT skills like spreadsheet, word processing, basic data manipulation skills that most Computer Science and Engineering graduates are expected to have.

But most of the categories are "soft" skills that may not be easily gained in most Computer Science and Engineering programs that we are aware of. "Communication" skills, at the top of list of soft skills, are the category that a majority (30.2 percent) of advertisements requires. Following these are "Analytical" skills (11.1 percent of all jobs), "Project Management" (4.6 percent) and "Presentation" skills (3.9 percent). The requirement of communication skills is typically qualified by the adjective "excellent" in an advertisement. This could suggest that the employers find these skills weak in Computer Science and Engineering graduates. About one in twenty five advertisements require "strong presentation skills" (3.9 percent) to the extent these advertisements specifically list these skills beyond the usual "excellent communication skills." One in fifty (1.6 percent) advertisements emphasizes "Teamwork," while 3.9 percent mention the ability to "Work Independently" (Figure 3).

Computer Science and Engineering programs could do more to teach these skills just as MBA programs help their students gain soft skills through team exercises, competitive exercises, group discussions, presentations and case analysis [5].

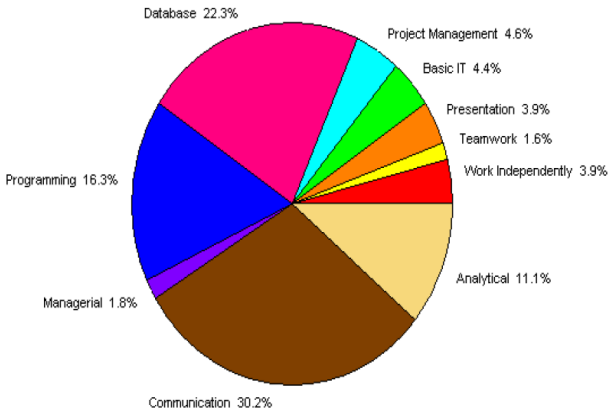


Figure 3 . Skills requirements

3.3. Job Location

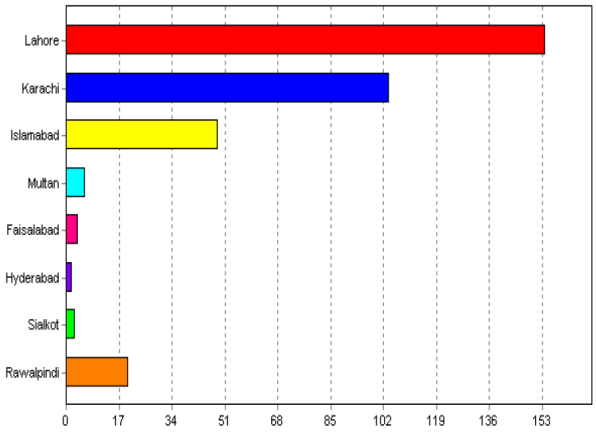
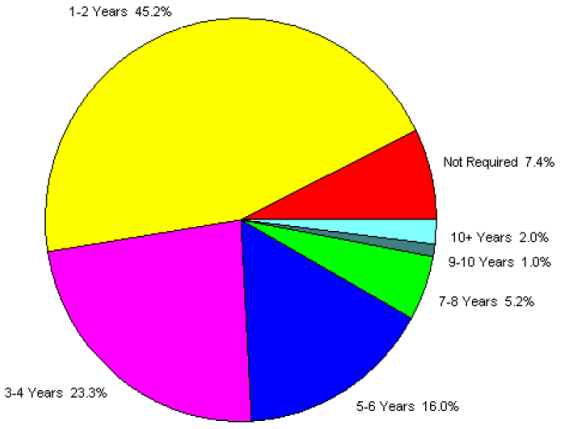


Figure 4 . Job location

As shown in Figure 4 most of the jobs were Lahore based followed by Karachi and Islamabad. This analysis also points to the fact that the majority of the companies are based in Lahore, Karachi and Islamabad, the big cities of Pakistan. This is the reason why these cities are most populated. Most of the students migrate to these cities after graduation and search for jobs. Students can use our analysis to move to a nearby city and to maximize their chance to get jobs after graduation. Universities may use our analysis to talk to these companies and help their students in getting internships in these companies prior to their graduation. This way they can prepare students for perspective job market and so students can easily get into these companies after graduation.

3.4. Experience Required

As shown in Figure 5 (a) about 50 percent of the jobs require one to two years of experience and about a quarter of the total jobs require three to four years of experience. Only 7.4 percent of jobs require no experience. This clearly points indicates that experience matters a lot to get into a good job. Students should realize that it is very difficult to find a highly paid job without having any experience. They should gain experience by working in their summer holidays as an internee with companies. This way they have a greater probability to get employed soon after their graduation. Universities should also focus on industrial training apart from teaching. They should make it an integral part of their curriculum. They should make at least five to six months internship compulsory as partial fulfillment of degree. University-industry partnership is very much important and universities should try to establish linkages with industries to get their students employed.



(a)

(b)

Figure 5 . Experience required

Indicated in Figure 5 (b) most of the jobs advertised are at experienced (non-manger) level. Very few jobs are at entry level. The experienced level jobs lead entry level jobs by quite a large distance. This makes experience very important to get into a good job.

3.5. Job Category

As apparent from Figure 6 about half of the job advertisements/companies are looking for Engineers. This is why Engineering is the most sought after profession in Pakistan. Students with highest grades go into Engineering and otherwise choose Computer Science. This analysis may help students to choose their career wisely after finishing their higher secondary school. We advise students not to choose a discipline blindly. They should get into a career by looking at the job market. This is a big problem with Pakistani students that they don't have any idea of the job market while they choose a career. After this publication we hope that students make an educated decision about their career.

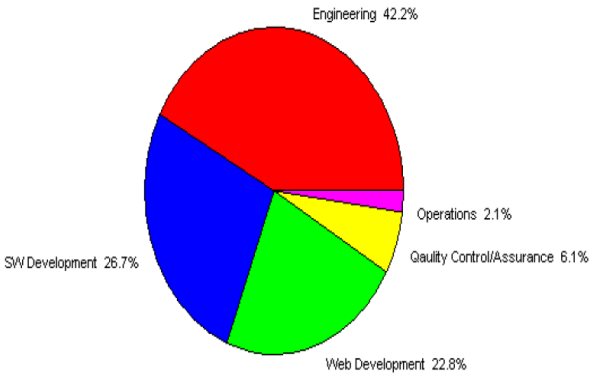


Figure 6 . Job category

3.6. Salary Range

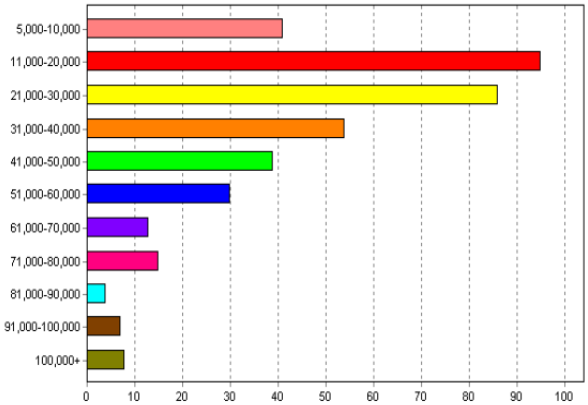


Figure 7 . Salary range

Most of the companies offer salaries between 10,000 PKR to 50,000 PKR per month and very few companies offers 100,000+ PKR. This analysis is also an indicator of the monthly income of Engineering/Computer Science graduates in Pakistan. This analysis may be used by multi-national companies to hire the services of Pakistanis graduates as labor is quite cheap here compared to other parts of the world (see Figure 7). Looking at market salaries Pakistan could be a better venue for outsourcing.

3.7. Job Type

Figure 8 compares different job types. As one can see most of the jobs are of type permanent and full-time. This analysis also shows the nature of people in Pakistan. Most people in Pakistan prefer to have a permanent job. They want to have job security even if they get a bit low salary. There are very few people doing freelancing. This may be due to the fact that most people are unaware of freelancing. There is a need to develop this awareness about freelancing.

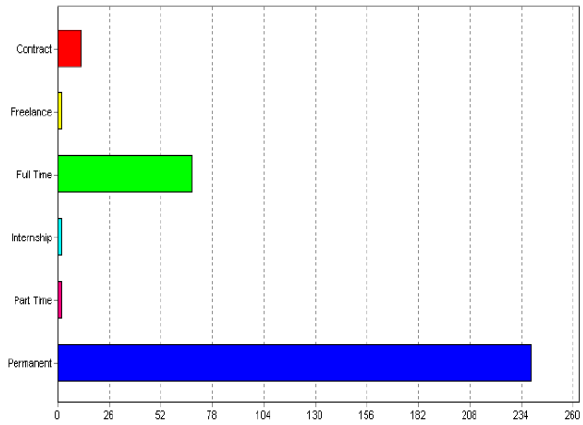


Figure 8 . Job type

3.8. Jobs by Gender

There is a common perception that there is a male dominant society in Pakistan. But the analysis shows that the situation is changing now. Most of jobs that were analyzed didn't mention any specific gender and were open to both male and female. But the condition is still a bit discouraging. The number of the jobs looking for male applicants only is far more than the number of jobs seeking only female applicants. As one can see in Figure 9 the number of jobs in which only female was sought is insignificant compared to jobs advertised for male applicants only.

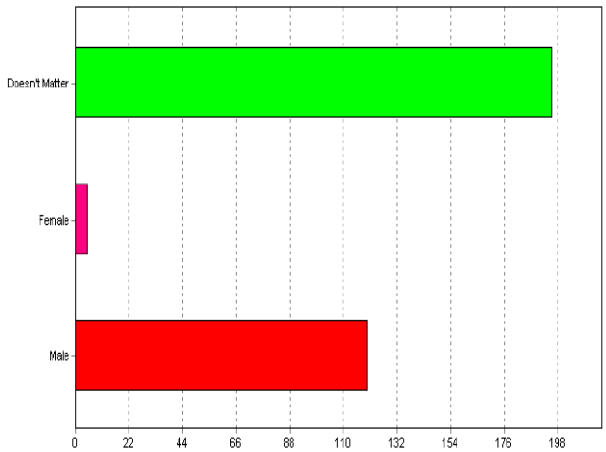


Figure 9 . Jobs by gender

3.9. Jobs by Date

The analysis (Figure 10) shows that more than 50 percent of the jobs were advertised in May/June. This is the time when the majority of students in Pakistan complete their degree. This analysis can be used by universities in Pakistan to make sure that their students graduate within this period. This way their students can easily be employed and otherwise they will need to wait for the next peak time to arrive. Most of the

universities don't take into consideration the job market when they design their academic calendar. As result, their students face tremendous difficulties in finding jobs when they finish their studies in the wrong time.

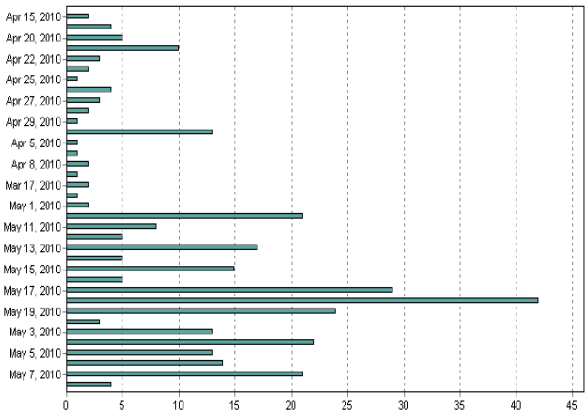


Figure 10 . Jobs by date

3.10. Number of Positions Advertised

The analysis shows that majority of the jobs or in fact almost all jobs that we analyzed required one to two persons. This clearly is an indicator of the unemployment and job crisis in Pakistan. According to a rough estimate approximately 20,000 Engineering/Computer Science students graduate every year in Pakistan. With so much production and so less consumption majority of the graduates find it really difficult to find a job. Government agencies can use our analysis to either restrict the production or increase consumption. There is a greater need to balance the equation.

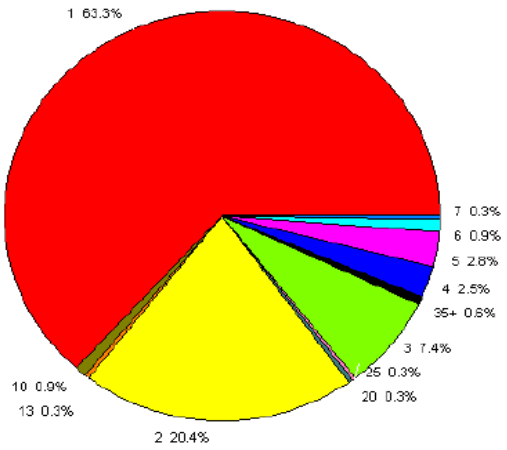


Figure 11 . Number of positions advertised

4. Summary & Conclusions

The results of text analysis of 1000 Computer Science & Engineering job advertisements, interviews and survey are published in the belief that these results would be useful for Computer Science and Engineering program directors, students, instructors and employers. The analysis indicates the following:

- The overall market for Computer Science and Engineering is strong. The number of advertisements in [Rozee.Pk] and [BrightSpyre.COM] that came up for different locations was 500+ and most of the jobs were from Software and Web development companies. Many were at experienced (Non-Manager) followed by Entry-level, but very few were at Manager and Executive levels.
- Demand for masters is smaller than that for graduates of bachelors. This may mean that, at least in Pakistan, students may be less likely to pursue master's degrees after bachelor's degree unless they are specifically interested in academia.
- Students in both graduate and undergraduate programs must try to develop the skills we outlined earlier if they want jobs in industry. Programming skills are quite important. C++ and Java are the most common requirement among job advertisements that require programming.

Going back to the goal, for students applying for jobs, especially online, at the very least the above should help them figure out what to mention and emphasize in their resume. Students starting their programs could use this analysis to figure out how they should supplement their program-related learning with additional skills gained on their own. For educators, the analysis may provide a discussion point in terms of redesigning their programs. They could also use the above to redesign homework by requiring students to work in groups and to use spreadsheets and/or database software. Presentations (not the same as PowerPoint slide shows) should be a norm for a class. Employers could also see that other employers are asking for similar things and use their links with universities to lean on Computer Science and Engineering programs.

Stemming from these results, it is believed that plenty of opportunities exist for growing the Computer Science and Engineering field and for placing our graduates in good jobs in industry and government. But also there is a need to actively consider the requirements of employers in optimally allocating our resources to what we teach and how we teach it.

Now we come to applications of our idea. This idea is new and so far Text Mining techniques are not reported in literature to be used for analysis of job advertisements. Our analysis is automatic as compared to other statistical approaches reported in the literature once data is given in required format. Our idea can be adopted to analyze different trends in other fields of life. For instance marketing people can use our idea to see customer behavior and accordingly design their marketing policies. Similarly our idea can be extended to examine the overall job market in a country/region. The same can be used for teacher and course evaluation in universities. One can do qualitative analysis once students' feedback is available.

5. Acknowledgement

The authors would like to acknowledge the support of the Electrical Engineering Department and King Fahd University of Petroleum & Minerals, Dhahran, Saudi Arabia. The idea of analyzing job advertisements comes from "*What Industry Wants from OR/MS Graduates"* by Man Mohan S. Sodhi *(m.sodhi@city.ac.uk)*, who heads the Operations Management and Quantitative Methods area at Cass Business School, City University London. For this the authors do acknowledge his original ideas and contributions.

6. References

[1] R. P.Schlee and K. R. Harich, **"**Knowledge and Skill Requirements for Marketing Jobs in the 21st Century,", Journal of Marketing Education, December 1, 2010; 32(3): 341 - 352.

[2] M. J. Liberatore and W. Luo, "The Analytics Movement: Implications for Operations Research,", Interfaces, July 1, 2010; 40(4): 313 - 324.

[3] R. J. Mooney and U. Y.Nahm. "Text Mining with Information Extraction,", Multilingualism and Electronic Language Management: Proceedings of the 4th International MIDP Colloquium, September 2003, Bloemfontein, South Africa

[4] M. A. Hearst. "What is text mining?" Oct. 2003.

[5] M. S. Sodhi, B. G. Son, and C. S. Tang, **"**ASP, The Art and Science of Practice: What Employers Demand from Applicants for MBA-Level Supply Chain Jobs and the Coverage of Supply Chain Topics in MBA Courses,**"**, Interfaces, November 1, 2008; 38(6): 469 - 484.