PREDICTING RELATIVE JOB PLACEMENT POTENTIALS AND MINING SKILL SETS BY ANALYZING ONLINE JOB ADS

by

Nevin OKAY

A thesis submitted to

the Graduate Institute of Sciences and Engineering

of

Meliksah University

in partial fulfillment of the requirements for the degree of

Master of Science

in

Computer Engineering

January 2016
Kayseri, Turkey

APPROVAL PAGE

This is to certify that I have read the thesis titled "Predicting Relative Job Placement Potentials And Mining Skill Sets By Analyzing Online Job Ads" by Nevin OKAY and that in my opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science in Electrical and Computer Engineering, the Graduate Institute of Science and Engineering, Melikşah University.

_____

Jan 22, 2016 Assoc..Prof. Ahmet UYAR
Supervisor

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

_____

Jan 22, 2016 Prof. Murat UZAM
Head of Department

Examining Committee Members

Title and Name		Approved
Assoc. Prof. Ahmet UYAR	Jan 22, 2016	_____
Assoc. Prof. Zeki YETGİN	Jan 22, 2016	_____
Asst. Prof. Kadir Aşkın PEKER	Jan 22, 2016	_____

It is approved that this thesis has been written in compliance with the formatting rules laid down by the Graduate Institute of Science and Engineering.

_____

Asst. Prof. M. Evren SOYLU
Director

Jan 2016

PREDICTING RELATIVE JOB PLACEMENT POTENTIALS AND MINING SKILL SETS BY ANALYZING ONLINE JOB ADS

Nevin OKAY

M.S. Thesis - Computer Engineering
January 2016

Supervisor: Assoc. Prof. Ahmet UYAR

ABSTRACT

Job ads contain quite valuable information. In particular, a large number of job ads in the web analyzed together offers the opportunity to reach significant statistics. In this study, online job ads relevant to engineering were analyzed using data mining techniques. In particular, classification and association analysis techniques were used. We retrieved 17.347 job ads from kariyer.net automatically with the help of a program we developed. We classify the job ads for five engineering departments: computer engineering, electrical engineering, industrial engineering, civil engineering, and mechanical engineering. We determined the total number of personnel advertised for each discipline. Using the total university placement quotas as the approximate number of graduates in Turkey for these departments, we calculated the relative job placement potential of each department. We used association analysis methods to examine the relationships of these departments in ads. In addition, we investigated most wanted technical skills for computer engineering.

This study contains valuable results for students who plan to select an engineering major in a university. In addition, it will be beneficial for computer engineering students who want to specialize on some subfields. Moreover, it will also be helpful for academics for curriculum development. It is the first study in Turkey by examining this many online job ads.

Keywords: Job ads, Data Mining, Job placement potential, Association analysis, Classification, Naïve Bayes, Chi-Square.

ONLINE İŞ İLANLARINI İNCELEYEREK GÖRECELİ İŞ BULMA POTANSİYELLERİNİ TAHMİN ETME VE YETENEK SETLERİNİ BELİRLEME

Nevin OKAY

Yüksek Lisans Tezi – Bilgisayar Mühendisliği
Ocak 2016

Tez Yöneticisi: Doç. Dr. Ahmet UYAR

ÖZ

İş ilanları oldukça değerli bilgiler içeriyor. Özellikle web'deki çok sayıda iş ilanı, birlikte analiz edilerek önemli istatistiklere ulaşma imkanı sunuyor. Bu çalışmada mühendislikle ilgili online iş ilanları, veri madenciliği teknikleri kullanılarak incelendi. Özellikle sınıflandırma ve birliktelik analizi teknikleri kullanıldı. Araştırma kapsamında yazdığımız bir program yardımıyla kariyer. net sitesinden 17,347 ilanı otomatik olarak çektik. İlanları bilgisayar mühendisliği, elektrik-elektronik mühendisliği, endüstri mühendisliği, inşaat mühendisliği, makine mühendisliği olarak 5 mühendislik dalına göre sınıflandırdık. Her bir sınıf için toplam personel sayılarını bulduk. Her bir ilanın kaç bölümle ilişkili olduğunu bulduk. Üniversitelerin bu bölümler için belirlediği toplam kontenjan sayılarını da kullanarak her bir bölümün göreceli iş bulma potansiyelini hesapladık. İlanların kapsadığı bölümlerin birbirleriyle ne kadar ilgili olduklarına birliktelik analizi metotlarını kullanarak karar verdik. Ayrıca bilgisayar mühendisliği için ilanlarda en çok aranan teknik yetenekleri ortaya çıkararak gruplandırdık.

Bu çalışma, eğitimciler, öğrenciler özellikle mühendislik tercihi yapacak üniversite adayı öğrenciler ve konunun diğer paydaşları için ilginç sonuçlar içeriyor. Türkiye'de bu alanda, bu boyutta yapılmış bir çalışma bulunmuyor. Veri analizi için kullandığımız yöntemler de sonraki çalışmalar için yol gösterici olabilir.

Anahtar Kelimeler: İş ilanları, veri madenciliği, iş bulma potansiyeli, birliktelik analizi, sınıflandırma, Naïve Bayes, Chi-Square.

DEDICATION

To my mother

ACKNOWLEDGEMENT

Thanks be to Almighty Allah for making me see the completion of this thesis project successfully.

I would like to express my gratitude to my supervisor Assoc. Prof. Ahmet UYAR.

My special thanks go to Asst. Prof. Hasan KİTAPÇI and Asst. Prof. Kadir Aşkın PEKER.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Technology is developing rapidly. When we compare today and a few years ago, we can easily see the difference. Changes in living conditions verify this rapid growth. Especially developments in information technology are fastest. For instance, in 2000, 5% of the population was using mobile phone, in 2014 this rate increased to cover all the population [1].

Rapid development of technology also makes it difficult to adapt to new technology. New products require new software to be developed quickly. The companies in this field need to watch closely the new technology for increasing their share in the market. Also the employees in this area should harmonize their skills with new technology. They need to renew their technical knowledge and skills based on changing technology. On the other hand curriculums of educational institutions must be updated to prepare the students to ever changing market frequently. Moreover, students should know the demanded skills by the market to better prepare themselves to the work life.

Job ads provide valuable data about the required skills by employers. There are many online job ad web sites. They provide detailed job descriptions and required skill sets. They host thousands of online job ads. These ads can provide very valuable information if collected and analyzed correctly.

It is not easy to find out clear and verifiable statistics for all the stakeholders of the topics. With emerging technologies, employers continually change the competencies of the staff they are looking for. Job descriptions are also changing. All these rapid changes complicate the formation of standards. The stakeholders are wondering that which capabilities new technology requires. It is not possible that the technological development stop and it is

quite difficult to predict the next step in the technological development. Despite all this uncertainty everyone needs a method to produce clear results with verifiable mechanisms.

1.2 MOTIVATION

Today unemployment is one of the most important problems. Previously it was easy to find a job after graduating from university. But now there are many unemployed graduates. Students attend the university to find a good job easier. And when they graduated they should have gained all of the capabilities, which that field required. But mostly education at university is not sufficient. Because technology is developing rapidly and universities cannot keep up with this development. Students often graduate with outdated skills. Many graduates are taking additional trainings to be able to catch technology.

On the other hand, employers also complain about not finding qualified personnel for the advertised jobs. Human resource professionals emphasize that universities must act in cooperation with industry. However, this cooperation does not take place at the desired level.

Analysis of the job ads could contribute positively to the solution of this problem. Job advertisements provide information about employers' expectations of potential employees and other information about evolving skill sets required in the technology workforce [2]. Analysis of the job ads may provide satisfactory solutions to the problems described above. In particular, the spread of online job postings and sites that have specialized in this field can provide verifiable and reusable solution. But it is difficult collecting and analyzing large amounts of data. Especially in the 80s and 90s, studies examined ads in the order of hundreds of ads. But this amount is not sufficient for robust conclusions. In addition, some of those studies used newspaper ads only, the others used a small number of online job ads. We couldn't see until 2013, a study using a sufficient number of job ads in this area. Most previous studies collected data and analyzed them manually.

However, thanks to the developments in Information Technology, we are able to manage very large amounts of data. Especially with the development of data mining and machine learning technology, we can find out significant results from stacks of meaningless data without needing a database. We can classify the data or group according to similarities, automatically. We can automatically analyze and summarize large stores of data to discover patterns and trends that go beyond simple analysis. But we saw very few studies in this field using the data mining techniques to collect and analyze data.

In this study, for the extraction of the desired results, we applied data mining methods. We collected about 18.000 online job ads from kariyer.net which is one of the Web sites in Turkey that has thousands job ads. We used text classification and associations analysis techniques for data analysis. We developed a method for collecting and analyzing data.

1.3 OBJECTIVE OF THE STUDY

- **Predicting the relative job placements potentials for top engineering graduates**

We want to determine the relative job placements potentials for top engineering graduates. For this, we first classify the job ads for each engineering discipline and calculate the total number of ads for each engineering department. Then, using the total university placement quotas as the approximate number of graduates in Turkey, we calculated the relative job placement potential of each department.

This information helps students, especially high school students. Because high school graduates does not know the job placement possibilities for engineering majors when selecting majors. They will find the outcomes of this study useful for this reason.

- **Discovering associations among engineering departments with the other departments:**

We analyze the co-occurrence of engineering names in ads and perform association analysis among engineering departments. For this, we try to find the number of ads which an engineering department co-occur with the other engineering departments. Then we try to find associations among engineering departments with the other departments. For this, we try to find the number of ads which an engineering department occur together with the other departments. So, we can find which departments are closer to the others.

- **Finding the most wanted skills for computer engineering:**

We try to find the most wanted skills for computer engineers. For this, we first extract job skill terms required for Computer Engineers from ads. Then we identify sub areas for computer engineering by grouping these skills. We try to find out how many ads there are in each sub-area. And this information helps educators and students at universities. Because the skills identified in this research can have important implications for the students in their selection of elective courses and when choosing a track for specialization. Educators will also

find the outcomes of this study useful for the design and development of new curricula that can prepare students for the job market.

- **Finding popularity of engineering disciplines among university candidates:**

Another objective of this study is to calculate the popularity of engineering fields among the university candidates. We use the base scores of students that prefer these programs from 10 universities in Turkey. We compare the popularity of departments to the relative job placement potentials.

Overall, our main goal is to develop a systematic method, which is reliable and repeatable. This is important because with the advancement of technology everything is changing in the industry. So, the results of this study should be repeatable if needed.

1.3.1 Contribution of the Study

- We have not seen any study calculating the employment potentials for Engineers. There are a few studies and all studies have been made to find the most wanted skills for Computer science graduates.
- Skill set analysis shows the demanded skills in the current job market. It should be helpful for academics to renew the curriculums and students to select the tracks in their majors.
- This is the first study for the job market in Turkey.

1.5 THESIS ORGANIZATION

In the second Chapter related works are given. In Chapter 3, the data collection steps are explained. In Chapter 4, the analysis of ads having engineering names explicitly is given. In Chapter 5, the analysis of ads not having explicit engineering names is provided. In Chapters 6, association analysis results are presented. Chapter 7 contains findings about the most wanted skills. Chapter 8 contains the conclusions and limitations..

CHAPTER 2

RELATED WORKS

2.1 RELATED WORKS

Starting in the late 80's, a few studies began to analyze newspaper advertisements. In the early 2000s, we see that online job postings provide data for researchers. During these years, there are some researches that used newspaper and online ads together. In addition to job postings, the questionnaires are also used to gather data in some studies.

Many studies in literature used a small amount of data. The main objective of these studies is to determine the set of most sought-after skills in the market. In addition, in some studies they tried to determine the most popular sub areas in computer engineering field. Furthermore, skills set analyses in job ads are used update the curriculums in some institutions. A common characteristic of these studies is that all have analyzed the ads for the computer industry. This is because may be, this sector most affected from the technological development.

Table 2.1 shows the summary of related works.

Ref	Year	Data	Methodology	Objective of the study
[2]	2009	241 online job ads	Manual examination	To examine IT job skills
[4],[3]	2014	Gather data for years 2005 –2014 from Dice.com	Web data mining(Keyword indexing)	To understand computer programming job trends
[5]	2010	209,655 unique job advertisements.	Web content mining (cluster analysis)	To determine most popular skill sets and most popular sub areas.

[6]	2003	300 online job ads	Manual examination	To identify technical skill categories
[7]	2006	200 online job ads from Monster.com	Manual examination	To identify the types of skills required for IT professionals
[8]	2007	250 jobs posted on Monster.com	Cluster analysis	To reveal basic skill groups & Curriculum Design
[9]	2011	131 newspaper ads	Manual examination	To assess the types of skills required by employers in Botswana.

Table 2. 1: Summary of Related Works

Huang, et al [2] they selected 241 online job ads listed on Monster.com from April 2008 - June 2008. They examined IT job skills. They used manual examination. For their analysis, they documented the prospective job company, the job description, and the complete job requirements broken down into Soft and Technical skills. These 241 jobs in their sample resulted in 5 humanistic (creativity, analytical, problem solving, dependability, hobbies), 17 technical (OO Programming, database, networking, MS Office, MS OS, MS Excel, SQL, Java, C++, C#, .Net, VB, XML, HTML, LINUX/UNIX, Project Management, SDLC methodology), and 7 business skills (leadership, communication, writing, organization, professional demeanor, teamwork, ability to work under pressure).

These results are summarized in Table 2. 1:

Skill	Skill category	Times referenced
1. Communication	Humanistic	172
2. Writing	Humanistic	116
3. SQL	Technical	93
4. Ability to work in teams	Business	72
5. Organization	Humanistic	65
6. Java	Technical	63
7. Problem solving	Humanistic	62
8. Analytical	Humanistic	61
9. Database	Technical	61
10. Professional demeanor	Business	49
11. OO Programming Languages	Technical	44
12. Microsoft OS	Technical	42
13. .Net	Technical	42
14. Linux / UNIX	Technical	41
15. SDLC methodology	Technical	40
16. C	Technical	32
17. C#	Technical	32
18. Dependable	Humanistic	30
19. HTML	Technical	29
20. Ability to work under pressure	Business	26

Smith & Ali [3] [4]. They searched specific keywords related with programming languages/technologies on Dice.com from 2005 until 2014. For this study the job site Dice.com was selected due to a focus on technical jobs and being amenable to the web mining process. They used web data mining (keyword indexing to collecting the data) as method. Their goal is analyzing computer programming job trend and to understand computer languages/technologies job trend. The total graph in Figure 2.1 provides the job demand for programming languages.

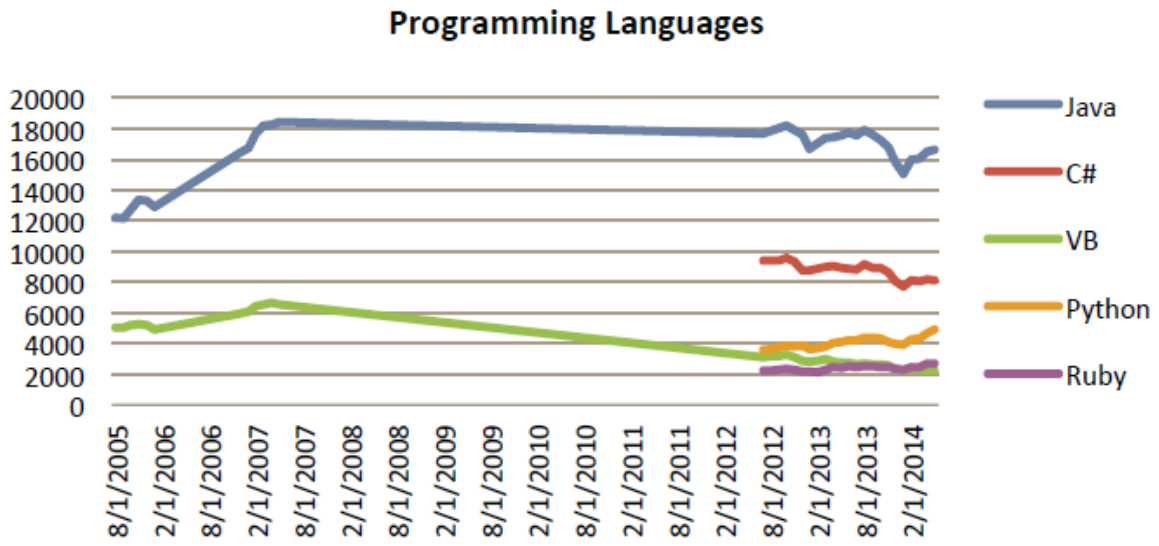

Figure 2. 1: General Purpose Programming Languages

Litecky et al.[5] developed software that systematically searched Monster.com, HotJobs.com, and SimplyHired.com daily between July 2007 and April 2008 for jobs requiring a degree in computer science, management information systems, computer information systems, and other computing programs. They extracted 209,655 unique job advertisements. Then they analyzed the data using cluster analysis.Cluster analysis revealed 20 job definitions, which they then verified using a manual review of 100 random job ads, with an overall successful classification rate of 91 percent. They performed a cluster analysis on the ads in two phases. Because they had no preconceived notions about the actual number of job types that the ads represented, they used hierarchical agglomerative clustering in the first step to identify 20 unique skill set clusters. They used this number as input to the second step: a *k*-means cluster analysis.

Table 2.3 shows the most frequently mentioned skills in computing job ads. Table 2.4 describes each cluster; Figure 2.2 shows the relative numbers of jobs from all job ads analyzed that were placed into each cluster.

Skill	Frequency (%)
Security	33.29
C/C++	28.69
SQL	27.57
Programming	26.08
Microsoft operating systems	23.18
Java/Java 2 Enterprise Edition/Java to Python	21.09
Leadership	20.10
Project management/planning/budgeting/scheduling	18.86
Software development	18.01
Oracle databases	17.19
Unix	17.15
Business strategy	17.06
Certification	14.68
Finance	13.98
XML	13.56
Generic databases	13.43
HTML/XHTML/DHTML	12.80
Open source operating systems	12.50
Marketing	12.47
JavaScript	12.10
Accounting	11.70
Microsoft databases	11.37
Object-oriented programming	11.16
.NET	10.55

Table 2. 3

Job definitions

Job title	Job description	Major skills required
Web programmers (6,187 / 3.0%)	Generic Web development using a variety of development platforms.	HTML*, JavaScript*, Java, XML, AJAX
MS Web developers (4,297 / 2.0%)	Web development specializing in Microsoft technologies.	C/C++†, ASP*, C#*, SQL*, HTML*, JavaScript*, XML, .NET, VB
MS Web application project analysts (3,843 / 1.8%)	Application development using primarily Microsoft technologies, including some system analysis.	C/C++†, SQL†, C#*, XML, MS Databases, .NET*, ASP, VB, OOP, SDLC
Java database Web developers (6,766 / 3.2%)	Web-based database application development using Java.	Java†, JSP, SQL*, MS, XML*, HTML, JavaScript, Oracle
Open source Web application developers (3,185 / 1.5%)	Web application development using open source technologies, GUIs, and back-end development.	HTML, open source/Unix operating systems, PHP, Java, JavaScript, Perl, SQL, open source databases
Java programmers (12,380 / 5.9%)	Programming position specializing in Java and Java-related programming.	Java†, programming, software development, OOP
MS developers (7,018 / 3.3%)	Traditional development specializing in Microsoft languages with high C# requirement.	C/C++†, C#†, .NET, object-oriented programming
Open source developers (8,258 / 3.9%)	Primarily a programming position working with many languages associated with the open source community.	C/C++, Java, open source/Unix operating systems†
C/C++ programmers (12,919 / 6.2%)	Programming specializing in C/C++. Few other major skill requirements.	C/C++, programming skills†
System-level C/C++ programmers (6,485 / 3.1%)	Specialized C/C++ programming, developing applications that interface at the operating system level.	C/C++†, programming skills*, security, operating systems*, TCP/IP, Perl, Java
Database developers (14,580 / 7.0%)	Working with SQL and different database systems. Moderate amounts of programming and system analysis.	SQL†, Oracle, MS and generic databases, programming skills
Java database application developers (7,210 / 3.4%)	Development of database applications in Java. Primarily focused on using Oracle and Unix.	Java†, Oracle*, SQL*, Unix*, Perl, XML
MS Visual Basic (VB) database application developers (6,849 / 3.3%)	Development of database applications primarily using VB, .NET, and ASP.	SQL†, Microsoft databases, Visual Basic, .NET, ASP
MS database application developers (6,975 / 3.3%)	Development of database applications using Microsoft technologies. Distinguished from MS VB Database Application Developer by requirement of C#, C/C++, SQL, and ERP skills.	C#†, C/C++†, SQL†, .Microsoft databases, NET, ASP
IT managers (26,656 / 12.7%)	Includes a variety of jobs, most of which include a leadership component as well as a high frequency of non-IT-oriented business skills.	Leadership, strategy, finance, marketing, accounting, telecom, CASE tools, SCM, BPR, ERP
System administrators (15,248 / 7.3%)	Administration of end-user computing systems and workstations (primarily MS operating systems) as well as networking and telecommunications.	MS operating systems†, security, certification, networking
Network administrators (7,982 / 3.8%)	Similar to system administrators but heavier emphasis on Unix, open source, Sun, and IBM operating systems. Special focus on networking multiple technologies.	Open source†/Microsoft/Unix/IBM operating systems, security, TCP/IP, Cisco, Perl
Database administrators (8,062 / 3.8%)	Works with the administrative component of databases. Oracle stands out as the dominant database management system (DBMS).	Oracle†, Unix*, SQL, databases, ERP, data warehousing, security
Security specialists (22,813 / 10.9%)	These positions all include some security aspect but are otherwise wide-ranging.	Security†, certification, leadership
Project analysts/ managers (21,942 / 10.5%)	Project management, often including a leadership or strategy component.	Project management planning†, budgeting†, scheduling†, leadership, strategy, certification, finance, ERP, responsibility

Table 2. 4: Job Definition

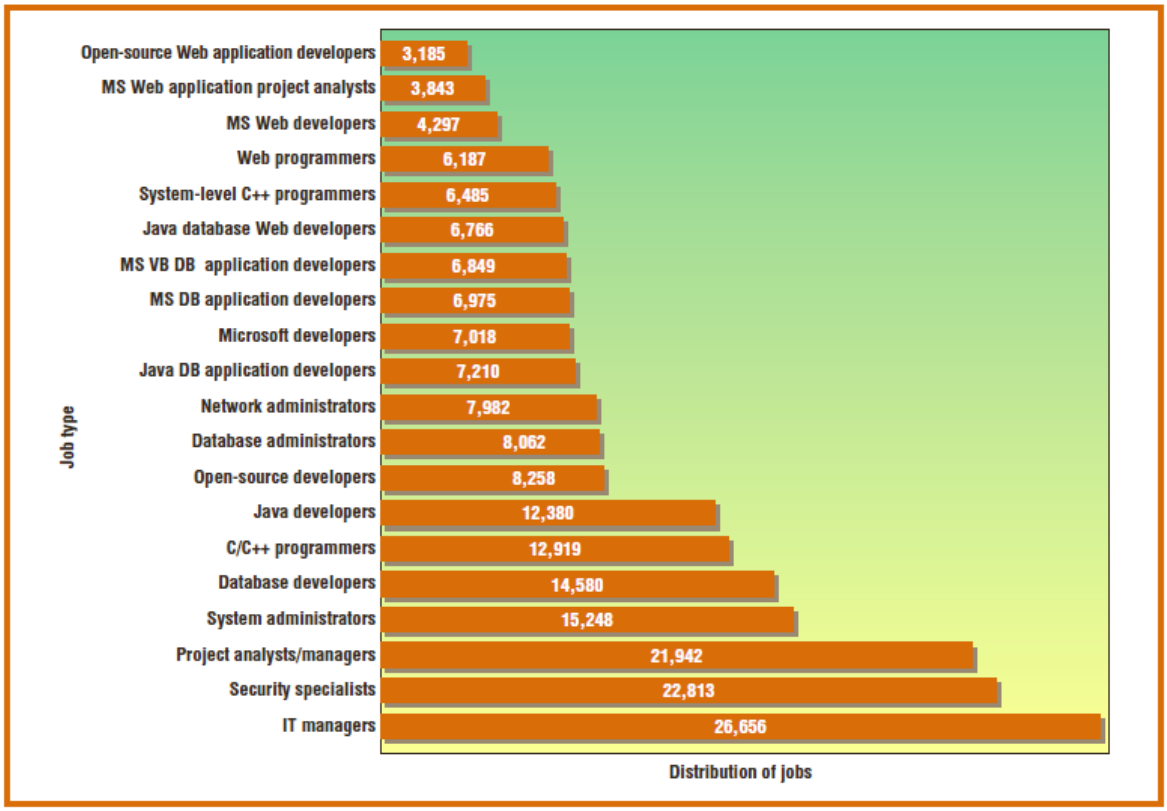

Figure 2. 2: Distribution of jobs by job type.

Koong et al [6] gathered the ad data from Monster.com and HotJobs.com required technical skills especially. They selected 300 online job ads. They analyzed data manually to identify technical skill categories and examining IT skill trend. They identified the following skill categories: programming languages (Java, C/C++, and Visual Basic were most frequent); website development (57% sought SQL and HTML
skills); databases (nearly 50% required Oracle); networks (only Windows NT or wide-area/local-area networks); and operating systems.

Liu et al [7] collected 200 online job ads from Monster.com. They analyzed data manually to identify the types of skills required of IT professionals. The job skills they find out in this study are divided into five categories. Each category is further divided into types of skills. The schematic model used for organizing the data set in this study is presented in Table 2.6:

Skill	Number	Percent
C	6	13.64%
C++/Visual C++	17	38.64%
Cobol	1	2.27%
Java	14	31.82%
Visual Basic	6	13.64%
Others	0	0.00%
Total	44	100.00%

Table 1: Skills in Programming Languages Category

Skill	Number	Percent
Java	25	49.02%
XML	11	21.57%
HTML	10	19.61%
Active Server Page	5	9.80%
Others	0	0.00%
Total	51	100.00%

Table 2: Skills in Web Development Category

Skill	Number	Percent
TCP/IP	4	57.14%
IPX	2	28.57%
HTTP	1	14.29%
Others	0	0.00%
Total	7	100.00%

Table 4: Skills in Networking Category

Skill	Number	Percent
DB2	9	29.03%
MS Access	1	3.23%
MS SQL Server	1	3.23%
Oracle	20	64.52%
Others	0	0.00%
Total	31	100.00%

Table 3: Skills in Database Category

Skill	Number	Percent
Linux	6	18.75%
Solaris	1	3.13%
Unix	18	56.25%
Windows	7	21.88%
Others	0	0.00%
Total	32	100.00%

Table 5: Skills in Operating System and Environments Category

Table 2. 5

G. Kent Webb. [8] selected 250 jobs posted monthly from February 2005 to February 2006 on Monster.com to reveal basic skill groups. He used cluster analysis for analyzing data. Clusters were determined automatically using Schwarz's Bayesian Criterion (BIC) and the log-likelihood distance measure. First he defined the skills then he defines which skills belong to which clusters.

Ayalew et al.[9] collected job ads from 7 major newspapers (Mmegi, Daily News, Gazette, Guardian, Midweek Sun, Sundays Standard, and Voice). The findings of this study have been used for the revision and development of curricula for undergraduate degree programmes at the Department of Computer Science, University of Botswana.

CHAPTER 3

CONSTUCTING THE AD COLLECTION

We first select five engineering departments. We explain the method that we used for selecting these majors. Second, we explain the ad collection method that we used and present the results.

3.1 SELECTION OF ENGINEERING DEPARTMENTS

At this stage, it was decided which engineering fields should be selected for this study. At the beginning of the study it was identified nine engineering fields. These are Computer Engineering, Electric-Electronic Engineering, Industrial Engineering, Civil Engineering, Mechanical Engineering, Environmental Engineering, Food Engineering, Chemical Engineering and Textile Engineering.

For the study we selected 5 engineering departments with the highest number of student enrolment capacity in Turkish universities. The Table 3.1 shows the total number of student quotas in nationwide student placement lists for the year of 2014. The top 5 engineering department from this list is selected.

Department	Total number of students
Mechanical Engineering	11248
Civil Engineering	10398
Electrical-Electronics Engineering	9896
Computer-Software Engineering	7939
Industrial Engineering	5290

Food Engineering	3925
Environmental Engineering	2860
Textile Engineering	489

Table 3. 1: The total number of student quotas

3.2 CONSTRUCTING THE AD COLLECTION

One of the main purposes of this study is to determine the relative employment potentials for the graduates of these 5 engineering departments. To achieve this goal, we decided to crawl and analyze the job ads on a popular job ad site in Turkey. We crawled the job ads from the most popular job ad site kariyer.net. It had more than 80,000 job ads as of March 2015. It covers ads from all regions and sectors of Turkey. We assume the ads on kariyer.net make a representative sample of all ads in Turkey. It is not focused on in any particular sector or region. In addition, the number of ads on the site is adequate for this study.

We want to crawl all job ads related to all 5 disciplines. In Kariyer.net website, ads can be listed by searching with keywords or filtering. For example, current ads, ads belong to a sector or all ads that contain a word can be listed with the help of the simple search engine on the site. So the ads on the site in order to be listed you need to determine a number of features. We also need to set these properties primarily. Because to extract the ads, we need to list them on the site. The easiest way for listing relevant ads on the site to search by keywords. This is more suitable for our purpose. Because we try to get all job ads relevant to the selected 5 engineering departments. We want to extract all of the relevant ads, and we want to leave outside irrelevant ads as much as possible. For this, we identified the search keywords and input the specific keywords by using search engine located on the website. We developed a java program to submit the queries and retrieve and save the result list. We saved all ad contents in text files on our hard drive.

3.2.1 Crawling Job Ads

We propose two steps ad crawling:

Retrieving all ads that explicitly contain the engineering major names: When we examine the ads we find two types of ads that relates to engineering. The first type of ads

explicitly contains the type of engineer they are looking for. The ads have the engineering names in the ads.

Some ads do not have the engineering names explicitly in ads. They specify the qualifications for the jobs only. However, a close examination of the required skills shows that these ads are primarily for the graduates of one of these departments. We consider these types of ads for engineering majors and we try to crawl them. The ads usually don't have the salary information in them. Therefore we cannot make a salary-based selection. We primarily examine the required skill sets in the ads and decide the type of person they are looking for.

3.2.1.1 We crawl both Turkish and English ads

In addition to Turkish ads, there are significant numbers of English ads also in kariyer.net. These ads also need to be collected separately and included in the classification. So we identified the keywords for English ads separately.

3.2.1.2 Ads Explicitly Contain The Engineering Major Names

For the first type of ads, we set a keyword to cover all engineering departments. For the Turkish ads we identified the keyword "mühendis", for English ads we identified the keyword "engineer". When we try these keywords on the search engine located on the site we get

- 8072 ads for Turkish
- 2068 ads for English

The word "mühendis" is the stem of the word for engineer in Turkish. There are many forms of this word in ads such as "mühendisi", "mühendisliği", "mühendislik", "mühendisler", "mühendisleri", etc. The search for the keyword "mühendis" returns all ads having these forms of the word. Therefore we use the keyword "mühendis" to retrieve all ads having explicit engineering major names.

Retrieved ads are all engineering majors. They are not limited to 5 selected majors. We initially try to retrieve all ads, and then we will determine the ad sets for each discipline later on.

3.2.1.3 Ads Not Explicitly Contain The Engineering Major Names

To query the second type of ads, we determine a set of skills as keywords for all engineering disciplines. We try to determine all relevant skills for each engineering discipline. We followed a multi step method to determine the skill sets for each discipline:

At the first step, candidate keywords have been identified manually for each Engineering Department. While doing this, we examined the content of ads and consulted with experts from each engineering department.

At the second step, we ignored some of the selected skill keywords based on the following criteria.

Rarely used skills are ignored. Some skills are rarely used in ad contents. We ignored those skills, since they don't help retrieving many ads. We randomly selected 100 ads from each disciple. We counted the occurrences of skill keywords on these ads. We ignored the skills that occur in less than 4 ads among these 100 ads.

Common skills are ignored. Some skills are very general and they retrieve many nonrelevant ads. For example the terms "web" or "yazılım" (software) are very general skill keywords. These types of keywords are ignored in this step. We submitted every keyword to the search engine of kariyer.net website. We examined the %10 of returned ads. We ignored the skills that returned more than %90 nonrelevant ads.

Duplicate skills are ignored. Some of the skill keywords are always used together. We manually examined the ads and determined such relationships. For example all ads having the keyword ABAP has also had the keyword SAP. Therefore, we ignored ABAP.

Computer Engineering	yazılım, web, programlama, android, ios, ağ, sql, oracle
Civil Engineering	msProject, sap2000, etabs, ohsas18001,midas, kesin hesap, isg, iso14001, ataşman, çelik konstrüksiyon, CAD, çelik dizayn, çelik yapı, yapı güvenliği
Electrical & Electronics Engineering	PCB, HMI, FPGA, baskı devre, eplan, servo motor, mikro dalga, görüntü işleme, biomedikal, MPLAB, matlab, assembly, netcad, mikro işlemci, mikro denetleyici, devre tasarımı, kontrol sistemleri.
Industrial Engineering	ABAP, simülasyon, arena, pro-model, risk yönetimi, raporlama, kurumsal kaynak yazılımı, netsis

Mechanical Engineering	asbuilt, shopDrawing, nastran,proengineering, makine-imalat, iklimlendirme, yapı denetimi, muhasebe, ısıl işlem, talaşlı imalat, bakım-onarım, mekanik sistem, doğalgaz, asansör denetim, unigraphics, kaynakçı, ARGE

Table 3. 2: Eliminated Keywords For Turkish Ads

We followed the steps outlined above for Turkish skill sets only. To determine the English skill sets, we simply translated the Turkish skill names to English. Table 3.3 shows the determined keywords for each Engineering Department. It has both Turkish and English keywords. Some keywords have only one form such as “php” or “Java”.

Computer Engineering

Skill Term	Number of ads on the site	Number of occurrences in randomly selected 100 ads
Database	353	
Yazılım uzmanı	944	28
Java	777	36
.net	469	21
Php	186	4
Mobil uygulama	471	12
Network	940	10
Veritabanı	430	22
Mobile application	137	
Software specialist	127	
C++	188	9
C#	533	27

	Skill Term	Number of ads on the site	Number of occurrences in randomly selected 100 ads
Civil Engineering	**Autocad**	817	33
	Primavera	133	18
	Metraj	359	22
	Şantiye şefi	264	21
	Hakediş	389	29
	İş güvenliği	366	9
	Site maneger	172	

	Skill Term	Number of ads on the site	Number of occurrences in randomly selected 100 ads
Electrical & Electronics Engineering	**Haberleşme**	288	9
	PLC	368	13
	Gömülü yazılım	49	4
	Gömülü sistem	46	4
	Scada	117	5
	Alçak gerilim	73	4
	Automation systems	63	
	Aydınlatma	228	4
	Embedded software	38	
	Embedded systems	40	
	Enerji uzmanı	323	18
	Orta gerilim	64	6
	Otomasyon sistemleri	346	5
	Yüksek gerilim	71	5

	Skill Term	Number of ads on the site	Number of occurrences in randomly selected 100 ads
Industrial Engineering	**SAP**	1755	27
	Satın alma	823	8
	ERP	1044	17
	Human resources	166	
	İnsan kaynakları	1613	14
	İş güvenliği	366	4
	Kalite kontrol	1393	19
	Performance management	547	
	Performans yönetimi	959	6
	Production planning	216	
	Purchasing	190	
	Quality control	320	
	Üretim planlama	1245	13

	Skill Term	Number of ads on the site	Number of occurrences in randomly selected 100 ads
Mechanical Engineering	**Autocad**	817	29
	Solidworks	289	8
	İş güvenliği	366	4
	Catia	147	4
	CNC	329	4
	Enerji uzmanı	323	7
	Kalite-kontrol	893	4
	Satın alma	523	4
	Purchasing	190	
	Quality control	320	

Table 3. 3: Determined Keywords For Turkish & English Ads

Third column of Table 3.3 shows the total number of ads for each skill. We downloaded these ads to hard drive and saved as text files. We saved the ads for each skill sets in a separate folder. For example for computer engineering, we retrieved all ads using skill terms which we determined for computer engineering, saved each as a text file and we collected these text files in a "computer engineering folder". We collected all computer engineering ads in the same folder therefore we didn't retrieve the same ad for two different skill terms. But all engineering departments have separate folders hence we retrieved the same ad for two or more different engineering departments. Table 3.4 shows the total number of ads for each engineering department extracted with these skill terms:

Mechanical Engineering	6096
Industrial Engineering	7853
Electrical-Electronic Engineering	1577
Civil Engineering	3643
Computer Engineering	3051
Total	22220

Table 3. 4: The total number of ads extracted with these skill terms

At this section we extracted 22220 ads from the site but these ads are not unique, an ad may be extracted for more than one department. However, in this data set there are ads that we retrieved before with the keyword "mühendis" or "engineer". We should remove these ads we don't need re-classified these ads. The total number of ads for each engineering department after removed these ads shown in Table 3.5:

Mechanical Engineering	4090
Industrial Engineering	5671
Electrical-Electronic Engineering	920
Civil Engineering	2202
Computer Engineering	1955
Total	14838

Table 3. 5: The total number of ads after removing ads with explicit major names

And also this data set contains some ads, which accept applications from high school and vocational high schools. We removed these ads. We consider such ads not to be appropriate for engineers. In addition, this data set contains some ads, which accept

applications from the departments that remain outside the scope of the research. We removed these ads. The total numbers of remaining ads are shown in Table 3.6.

Mechanical Engineering	1651
Industrial Engineering	2356
Electrical-Electronic Engineering	301
Civil Engineering	814
Computer Engineering	1319
Total	6028

Table 3. 6: The total number of remaining ads

3.2.2 Automatic Retrieval of Ads

We developed a Java program to retrieve the ads from kariyer.net web site programmatically. The program submits the queries to the website and retrieves the result lists. It uses an HTML parser, which creates a Document Object Model (DOM) tree. We used Java's parser libraries for parsing. We developed a content extraction algorithm for kariyer.net. As a rule a specially developed content extraction algorithm is required for each individual data source, because of the different and unique structures of websites. We used Jsoup library for implementation because of its implementation language is java besides it supports both cleaning and parsing HTML. In order to construct the DOM tree of the input web page correctly, HTML file needs to be well-formed. But most web pages are not well-formed documents. They contain invalid tag structure such as there is an opening tag with no corresponding closing tag and vice versa. Some HTML tags are nested in wrong order and also some tags are mixed up. Therefore these invalid tag structures are needed to be cleaned before processing them.

Extracting the content of ads was held in two phases. In the first phase, we submitted the keywords for listing the ads. Then we extracted the ad links. In the second phase, we connected to the website for the content of each ad.

3.3 CONCLUSION

In this chapter, we constructed the data set to be used in this study. First, we decide that which engineering departments covered by this study. We selected 5 engineering departments with the highest number of student enrolment capacity in Turkish universities. We have prepared a data set consisting of job ads which can be relevant to selected 5

engineering departments. Our goal is to find the employment potential of the five engineering departments separately. That's why we use job ads as data set. We constructed our data set with relevant job ads as much as possible. We crawled job ads from "Kariyer.net" via a java program.

We have constructed two different data set. We have constructed the first data set by retrieving all ads that explicitly contain the engineering major names. And we have constructed the second data set by retrieving all ads that don't contain the engineering major names explicitly.

CHAPTER 4

DATA ANALYSIS

By analyzing data we are going to reach two main results:

1. Possibility of job placements for each engineering department

   Possibility of job placements = number of job ads/ number of graduated students
   We used this formula for each engineering department separately.

   - To estimate number of job ads for each engineering department I should classify job ads using classification technique.
   - To estimate number of graduated students I use the 4-year-old student capacity of each engineering department in Turkish Universities.

2. Determining the sub areas of computer engineering

   - Determining the popularity of each sub area
   - Determining the key capabilities of each sub area

We are going to find the most wanted skills for computer engineers. For this, first, we extract job skill terms required for Computer Engineers from ads. Then we identify sub areas for computer engineering by grouping these skills. And we find out how many ads there are in each sub-area.

4.1 ANALYZING PROCESS

17,347 unique job ads, related with the research, have been extracted from the site Kariyer.net. However, these data do not have the same characteristics. Therefore, the analysis processing is performed in several steps using different techniques. Our main goal is classifying all advertisements have been collected, for this, we use different classification techniques because of different characteristics of ads. After finding total number of each engineering department, for computer engineering we identified subareas namely we again

analyzed the computer engineer ads separately and we found out the total number of each sub-area. So analyzing process contains 3 classification steps.

1. In the first step, advertisements obtained by using keywords "engineer" and "mühendis" were classified.
2. In the second step, we classified ads not explicitly contain the engineering major names.
3. In the final step, we classified the computer engineering ads in itself.

4.1.1 Calculating the Relative Potential of Job Placements With Ads containing Explicit Department Names

We retrieved 10,140 unique job ads that explicitly contain the keywords “mühendis” or “engineer”. As the first step of the analysis, we need to classify the ads based on the engineering disciplines. We need to determine the set of ads that are looking for the graduates of each engineering discipline.

Our analysis of ads on this section is based on the processing of explicit discipline names in job ads. For example, if an ad is looking for a computer engineer, it must explicitly state this in the ad by using an explicit form of the discipline name. The text “computer engineer”, “bilgisayar mühendisi” or a derived form of these two phrases must appear in the text. Similarly, when they are looking for the graduates of other engineering departments, they must explicitly provide the names of those disciplines in the ad text.

We use the keyword based search method to classify these ads. If an ad contains some derived form of engineering discipline name, we assume that this ad is looking for that type of engineer. Therefore, the ads have been classified by searching the names of engineering disciplines in the ad text.

4.1.1.1 The difficulty of classification:

- The most important problem is spelling mistakes: Spelling mistakes made in the engineering domain names used in this classification has affected the classification. To reduce the negative impact of errors made in this way, common misspellings in the engineering field names are included in the word list for using classification. For example, the word "makine" is misspelled as "makina" in a lot of ads. Therefore, the word "makina" was also considered correct word.

- The second problem the ad texts are the lack of a standard: The ad texts consist of 4 parts: ad title, job description, general qualifications and ad info. Graduation info (that should be considered graduates of engineering applications in which information) can take place in any part of the ad text. For this reason, the search was carried out in all the text.
- The third problem, ads include some information about the company that is placing the ad. Although these parts of the text are not related to our research, it is difficult to remove these sections from the ad texts, because these sections can be located anywhere in the ad text. The company information may introduce some errors when performing classification. For example, when a company in the construction sector is describing the work area, the text always contain the word "inşaat". However, such an ad may not search/seek a civil engineer. Therefore, to minimize the classification errors, engineering names aren't searched alone. A search was performed as described in the classification process section instead.
- The fourth problem is the similar to the third one. There is a section in ads about the business sector of the company. This information can sometimes be company based rather than ad based. This was also the reason for the incorrect classification for some class especially mechanical and civil engineering. For this we remove this section from the ad text before the classification.

4.1.1.2 Characteristic of classified ads:

- There are ads that accept applications from all the engineering departments without mentioning any particular department name: We omit these ads when performing classification. We cannot put these ads in any category based on the engineering discipline names.
- There are ads that accept applications from multiple departments: If an ad accepts the applications from multiple engineering fields, we proportionately relate this add to all those engineering disciplines. If an ad accepts applications from both computer engineers and electrical and electronics engineers, then this add is related with %50 with computer engineering and %50 with electrical and electronics engineering.
- Calculation the number of personnel: For this study, the total number of personnel in ads is important for each discipline. Total number of ads is not important. Therefore, we examined the ad contests and determined the number of personnel they are

looking for. For example, an ad accepts the admission of 3 engineering departments and also this ad seek 2 personnel. For this ad, we add 2/3 personnel for each of the 3 engineering departments. The number of personnel info is taking place in most of ads. For ads, which do not mention the number of personnel, the number of staff was accepted as one.

4.1.1.3 Classification Process:

Classification process was done entirely by keyword search method in this step. Turkish and English ads were classified separately. At this stage, 8072 Turkish ads were extracted by using the "mühendis" keyword and 2068 English ads were extracted by using the "engineer" keyword has been classified.

In this section, both we classified ads according to five engineering departments, and we calculated the total number of personnel for each engineering department. When calculating the total number of personnel for each class, we used personnel info located in ad text and we used university department info that indicated how many departments can apply. Personnel info is located in a specific place. So, using a code, it is possible to find this information in the text, and doing calculations, automatically. But also there are ads don't specified number of personnel. For this type of ads the number of personnel we assumed as one. The main problem in finding the number of personnel, there are some ads which the number of personnel info located more than one place in ad text. The program we wrote gives a warning in such a case. We fixed these type of ads manually. We deleted the number of personnel info in the wrong place so we provided to locate in just one place for all ads. Thus, we find the number of personnel for each ad automatically.

The number of university department info which we use when we calculated the total number of personnel isn't easy to find as the number of personnel. Because this info doesn't locate in specific place in ad text and doesn't use with a distinctive title. We use the keyword search method finding university department. Located in ad text, we need to find the names of all departments covered or uncovered by the research, so we use a two-step method.

4.1.1.3.1 Step One: Classifying Ads for 5 Engineering Departments

At the first step, we searched 5 engineering departments that we determined. This step is carried out by searching the engineering names in ad text. However, to minimize the classification errors caused by company advert located in ad text, the simple engineering names are not searched alone. Instead, words or phrases as shown in Table 4.1 is used to search. The words or phrases for English discipline names are shown in Table 4.2.

Computer Eng.	**Mechanical Eng.**		**Civil Eng.**	**Electrical-Electronics Eng.**		**Industrial Eng.**
bilgisayar mühendis	makine mühendis	makina mühendis	inşaat mühendis	elektrik mühendis	elektronik mühendis	endüstri mühendis
bilgisayar müh.	makine müh.	makina müh.	inşaat müh.	elektrik müh.	elektronik müh	endüstri müh.
bilgisayar/	makine/	makina/	inşaat/	elektrik/	elektronik /	endüstri/
bilgisayar-	makine-	makina-	inşaat-	elektrik-	elektronik -	endüstri-
bilgisayar ve	makine ve	makina ve	inşaat ve	elektrik ve	elektronik ve	endüstri ve
bilgisayar veya	makine veya	makina veya	inşaat veya	elektrik veya	elektronik veya	endüstri veya
bilgisayar,	makine,	makina,	inşaat,	elektrik,	elektronik ,	endüstri,
bilgisayar\	makine\	makina\	inşaat\	elektrik\	elektronik \	endüstri\

Table 4. 1: Words/phrases used in classification for each of the engineering fields in Turkish ads

Computer Engineering	**Mechanical Engineering**	**Civil Engineering**	**Electrical-Electronics Engineering**		**Industrial Engineering**
computer engineer	mechanical engineer	civil engineer	electrical engineer	electronics engineer	industrial engineering
computer/	mechanical/	civil/	electrical /	electronics/	industrial /

computer -	mechanical -	civil-	electrical -	electronics-	industrial -
computer and	mechanical and	civil and	electrical and	electronics and	industrial and
computer or	mechanical or	civil or	electrical or	electronics or	industrial or
computer,	mechanical,	civil,	electrical ,	electronics,	industrial,
computer \	mechanical \	civil\	electrical \	electronics\	industrial \

Table 4. 2: Words/phrases used in classification for each of the engineering field in English ads

This gradually improved the query list. We started with a few phrases and added more. After each iteration, we checked the classified ads. We tried to reduce the error rate by finding the causes of misclassifications. We set some classification rules to improve the system and to minimize the misclassification as follows:

1) To minimize the classification errors caused by sector section located in ads text, we find this section and remove the ads text automatically. Most of the incorrect classification stems from this section especially for mechanical and civil engineering. This section is mostly located in the same place in the text. However, this section can be in more than one place in some texts. We try not to make a change in the text that could negatively influence the classification as much as possible when removing this part from the text. So we just do this additional control for mechanical and civil engineering for Turkish ads and electrical-electronics engineering for English ads. Figure 4.1 is an example for this misclassification:

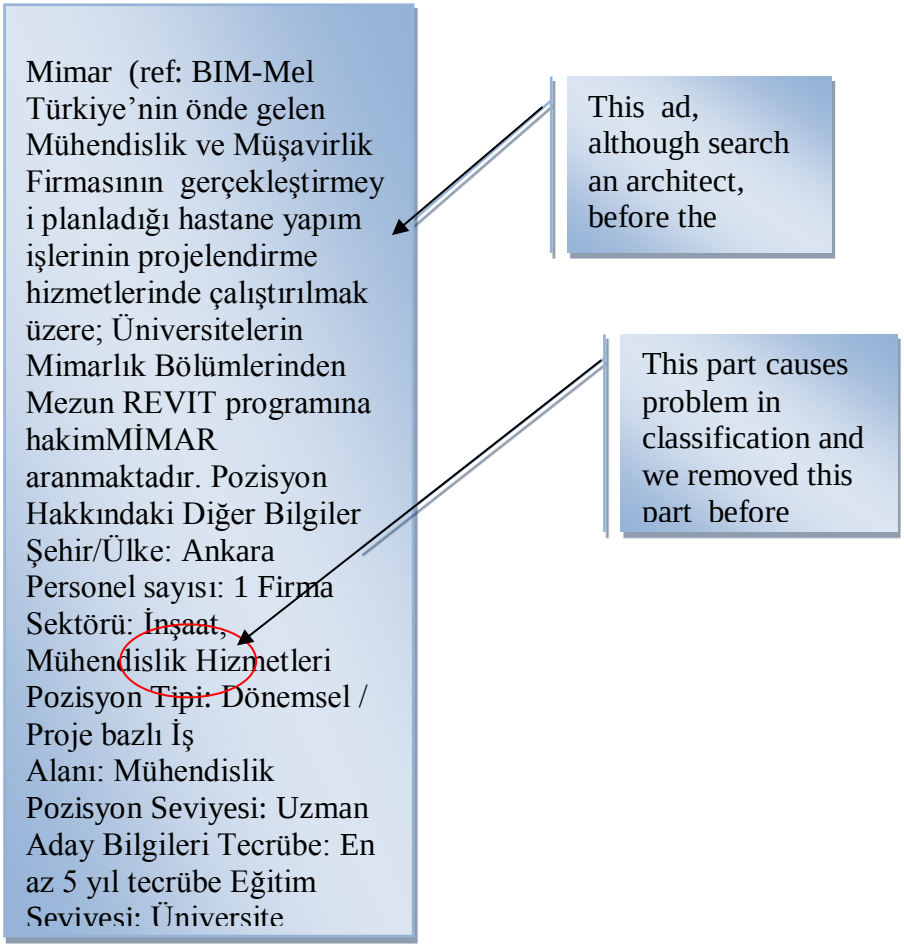

Figure 4. 1: The classification errors caused by sector section

2) There are some ads that don't have any relevant engineering department names. Also this type of ads cause classification errors. These ads can contain words using classification but these ads don't seek an engineer. So we try to differentiate this type of ads from the other ads. For this, we use educational level info placing the ad text. There is as a separate section contains this information in the texts. However, sometimes this section can take in more than one place in the text by including different information. Therefore, we search this information until end of text and as the educational level if not specified the level of education, which is expected to have an engineer, we omit these ads as seen in Figure 4.2 when performing classification.

3) For English computer engineering ads, sometimes it is used "computer science" instead of "computer engineer" for graduation information. So we use both "computer science" and "computer engineer" phrases for classification.

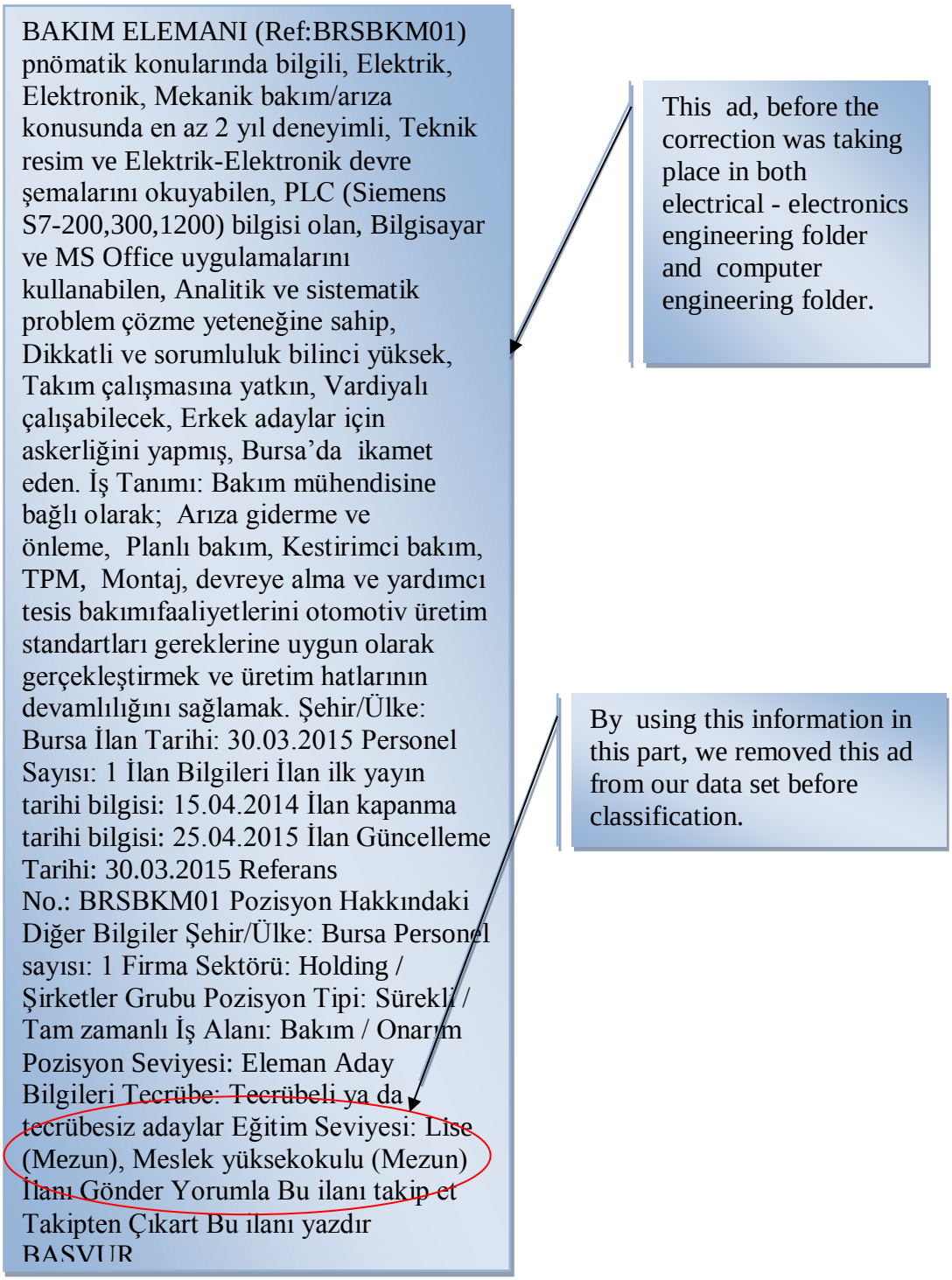

Figure 4. 2: This is an irrelevant ad with the research.

At this step, with this method, 8072 ads in Turkish and 2068 ads in English are classified and the results are shown in Table 4.3 and Table 4.4.

During the classification process, we put the ads that do not contain any of the 5 discipline names to the "other" category. This category has ads that accept applications from all the engineering departments, or accept applications from some engineering departments

that remain outside 5 disciplines, or non relevant for any engineering department but mistakenly retrieved.

Department	Total Number of Ads
Mechanical Engineering	1895
Electrical-Electronic Engineering	1484
Industrial Engineering	1467
Civil Engineering	946
Computer Engineering	871
Other	3038
Total	9701

Table 4. 3: Result for Turkish ads

Department	Total Number of Ads
Computer Engineering	364
Mechanical Engineering	442
Industrial Engineering	420
Electrical-Electronic Engineering	362
Civil Engineering	115
Other	829
Total	2532

Table 4. 4: Results For English ads:

The error rate of this classification method was identified by examining randomly selected 100 ads from each category. Table 5 shows the error rates for each engineering discipline. The error rates are all less than %5. This means that the ads that we classified as belonging to one of these disciplines are highly accurate. In addition, the error rates for the other category is less than %5. This means that at most %5 of ads are incorrectly placed in other category.

Department	Error Rate for Turkish Disciple Names (%)	Error Rate for English Disciple Names (%)
Mechanical Engineering	5	3

Electrical-Electronics Engineering	4	5
Civil Engineering	5	3
Industrial Engineering	0	0
Computer Engineering	2	0
Other	4	5

Table 4. 5: The Error Rates

We further examined the misclassified ads and determined their correct placements manually. Table 6 shows the results. Each row of the table shows 100 ads examined for each department. For example, second row shows that for 100 ads classified as mechanical engineering, 95 of them are correctly classified. 1 ad out of 5 incorrect classified ads belongs to Electrical-Electronics Engineering department and 4 ads belong to the other category.

The total of each column at the last row shows the total number of ads that belong to the corresponding department. This number is the result of manual classification. For example, for the 600 examined ads, there are 101 ads for mechanical engineering. The difference between this number and 100 shows the overall error in classification. Out of these 600 ads, manual classification determines 101 ads for mechanical engineering and algorithmic classification determines 100 ads. Therefore, we may conclude that the error for mechanical engineering classification is |101-100| = 1. Similarly the error for other departments can be calculated. They are all less than 2%. The error is higher in the other category though. It is 5%.

	Mechanical Eng	**Electrical-Electronics Eng**	**Civil Eng**	**Industrial Eng**	**Computer Eng**	**Other**	
Mechanical Engineering	95	1	0	0	0	4	100
Electrical-Electronics Engineering	2	96	1	0	0	1	100
Civil Engineering	1	0	95	0	0	4	100
Industrial Engineering	0	0	0	100	0	0	100
Computer Engineering	2	0	0	0	98	0	100
Other	1	1	2	0	0	96	100
Total for manual classification	101	98	98	100	98	105	

Table 4. 6: Manual examination results of incorrect classifications

Misspellings are the most important reason of errors. We have tried to eliminate all the errors we can generalize as much as possible by identifying as the classification rule. However, some classification errors caused by specific statements and company advertising continues to be a problem especially for mechanical, electrical and civil engineering fields.

Conclusion for step one: At this step, we grouped the ads according to engineering department names. So we found the total number of ads for each engineering department that covered by this research. By grouping ads in this way it will facilitate finding of meaningful relationship with the association analysis done at the second step. Because by grouping ads firstly, we can eliminate some irrelevant ads before the association analysis. Secondly, we are generating meaningful relatively small data sets from the total data. Due to these data are consistent within itself and the groups are compatible with the association rules that we are trying to find, association analysis algorithm can be implemented for each data group independently from each other. We expect this method makes it easier determination of the association rules.

4.1.1.3.2 Step Two: Determining other department names in ads

We have created a data set relevant with engineering departments. However, the most ads in our data set also relevant with some departments remaining outside the scope of our research and these ads text contain these departments name explicitly. For example, an ad look for industrial engineers can look for management engineer at the same time. Or an ad look for civil engineer can also look for an architect. We need to consider this when calculating the number of personnel for each engineering department. So, if an advertisement accepts the applications from multiple departments, whether relevant with engineering or not, a proportional calculation must be done as described above for calculating the number of personnel. For doing this calculation, we need to reveal the names of departments remaining outside the scope of our research from data set.

At this step, firstly we try to determine the university department names remaining outside the scope of research. Secondly, we estimate the total number of personnel for five engineering departments.

For determining the university department names, we use folders that we generated the previous step for the data set and we use the FPGrowth algorithm to discover the names. By this algorithm, for finding significant associations between words, first we have prepared the data with pre-processing. Data pre-processing consists of removing stop-words and non-alpha numeric characters from documents, elimination of some words according to word length and stemming process. And we also eliminate some words according to word type. Namely, we use only noun words for Turkish ads. This pre-process is performed separately for Turkish and English ads. Stop-word lists were created separately for ads in Turkish and English. For stemming operations "Zemberek" is used for the Turkish text and "Snowball Stemmer" is used for the English text. Word length 3 or less and word length 15 or more are eliminated. Thus a word vector for each document was obtained. We combined these word vectors in a text file. With data preprocessing, we achieved to obtain more acceptable size and less complex file. This file will be used in association analysis step is complex as little as possible will make the association file, after the analysis process we obtain, also less complex and examining this analysis file will be easier by hands to find significant associations.

4.1.1.3.2.1 Association Analysis Process: Fp Growth Algorithm

FP Growth algorithm is used to perform this analysis. FP Growth is one of the fastest data mining association algorithms. Association algorithms are used to frequent item set mining. FP Growth algorithm is commonly used for discovering frequently co-occurrent item sets.

FP-Growth works in a divide and conquer way. It requires two scans on the database. FP-Growth first computes a list of frequent items sorted by frequency in descending order during its first database scan. In its second scan, the database is compressed into a FP-tree. Then FP-Growth starts to mine the FP-tree for each item whose support is larger than minSup by recursively building its conditional FP-tree. The algorithm performs mining recursively on FP-tree. The problem of finding frequent item sets is converted to searching and constructing trees recursively.[10]

4.1.1.3.2.2 Finding Association Rules

As explained above, we used the FP Growth algorithm to define association rules. Although we have prepared the data with pre-processing and grouped the ads, at the first experiments we found too many rules to examine by hand. That's why we decided that we need to limit these rules. For this we followed the 3-step optimization process.

1) **minsup-maxsup parameters:** By using these parameters we limit the transaction data considerably. We have set these values for each group differently according to the groups' data sets. By setting minsup-maxsup values we're just getting the association rules within that range. For example, we determine as minsup = 50 and maxsup= 400, we have eliminate all the association rules which are larger than 400 or smaller than 50. minsup and maxsup values are indicated the frequencies of that association in the searched data set.
2) **Reference Words:** We have identified a reference word for 5 folders performed association analysis. This word must be found in all the association we look for. So, we decided that the most appropriate reference words are class names. For example, we have set the reference word: "computer" for ads we classified as "computer science" and collected in a folder. Or we have set the reference word: "endüstri" for ads we classified as "endüstri mühendisliği". In this way, for Turkish and English ads and for each engineering field we identified reference words separately. Thus, we have eliminated most of the association in the range of minsup-maxsup. In this way, we have found the association rules that both in the range of minsup-maxsup and the reference word have been found in. But there is a contradiction here. A reference word belongs a class exists in all ads belonging to that class. So the frequency of a reference word must be greater than the maxsup values. Then, the algorithm in this form can not find the association rules we want to find. We solved this problem by giving privilege to the reference words. So, even if the frequency values of these words are greater than maxsup, we allow them to be included in the association tree and we have achieved to get all association in the range of minsup-maxsup.
3) **Finally we limited the number of words in association:** We keep each association in an array and we assign a constant range to size of array. Size of array must be greater than 1 and less than 6. Thus, we have eliminate the association rules formed by one word, and more than 5 words. Because associations that are meaningful for us cannot be less than 2 and greater than 5.

As a result we get different association rules for each engineering department. By examining the associations we have determined university departments' names remaining outside of our research for five engineering disciplines separately. These department names we identified and their frequency values shown in Table 4.7.

1.	İşletme (Mühendisliği)+Management Engineering+ Business Administration	826
2.	Matematik Mühendisliği+Mathematical (Engineering)	597
3.	Mimarlık+Architecture	367
4.	Mekatronik +Mechatronics Engineering	262
5.	Malzeme ve Metalurji Mühendisliği+Material and Metallurgical Engineering	242
6.	Kimya Mühendisliği+Chemical Engineering	213
7.	İktisat + Economics	210
8.	Maliye+Finance	198
9.	Ekonomi+Economy	197
10.	Haberleşme Mühendisliği+Communications Engineering	195
11.	İstatistik+Statistics	155
12.	Harıta Mühendisliği+Geomatics Engineering	96
13.	Yönetim Bilişim Sistemleri+Management Information Systems-MIS	91
14.	Şehir ve Bölge Planlama+City and Regional Planning	47
15.	Psikoloji+Psychology	41
16.	Telekominikasyon+Telecommunication	38
17.	İç Mimarlık+Interior Architecture	37
18.	Çevre Mühendisliği+Environmental Engineering	34
19.	Tekstil Mühendisliği+Textile Engineering	31
20.	Biyomedikal+Biomedical	24
21.	Uçak Mühendisliği+Aeronautical Engineering	22
22.	Enformatik+Informatics	22
23.	Gıda Mühendisliği+Food Engineering	22
24.	Fizik Mühendisliği+Physics (Engineering)	21
25.	Muhasebe+Accounting	20

Table 4. 7: Identified department names and their frequency values.

The frequency values in Table 4.7 shows the number of times each department names occurs in ads with explicit department names for 5 disciplines. We require each department names to appear at least 20 times to include in this list. The values are combined values from Turkish and English ads. In total we determined 24 discipline names.

Although software engineering has a higher frequency value this table doesn't include this department name. Because software engineering quite similar with computer engineering and if an ad look for software engineer, we assume that this ad look for computer engineer.

And also because of similarity between electrical engineering and electrical electronic engineering if an ad look for electrical engineer, we assume that this ad look for electrical electronic engineer.

4.1.1.4 Estimate The Total Number of Personnel

Our program compute the total number of personnel for each engineering, works as follows:

I. Find the number of personnel data for an ad=perInfo;
II. Find how many university departments are relevant with this ad=relDep;
III. Proportion these two: $Prop_i$=perInfo/relDep;
IV. Add this ratio to the total number of staff of each engineering department that relevant with this ad and covered by the research:$Total_i += Prop_i$

With this method, we computed the total number of personnel for each engineering department are shown in Table 4.8. The overall accuracy of this system is 91%. We identified the error rate by examining the value of two variables (perInfo and relDep) during the runtime. While the program is running, first we randomly selected 100 ads and then we saved to a file the values of perInfo and relDep variables and ads text of this selected 100 ads. Thus, we determined the error rate by controlling this file. All errors caused by variable relDep.

Department	Total Number of Ads	Total Number of Personnel
Mechanical Engineering	1895	1407,7
Electrical-Electronic Engineering	1484	1124,7
Industrial Engineering	1467	785.02
Civil Engineering	946	749,9
Computer Engineering	871	632,3
Other	3038	
Total	9701	4669,7

Table 4. 8: Result for Turkish ads

Department	Total Number of Ads	Total Number of Personnel
Computer Engineering	364	263,7

Mechanical Engineering	442	313,6
Industrial Engineering	420	207,4
Electrical-Electronic Engineering	362	215,4
Civil Engineering	115	80,8
Other	829	
Total	2532	1080,9

Table 4. 9: Result for English ads

Table 4.10 shows the number of ads that accept applications from only one engineering department covered by research:

Department	Turkish Ads	English Ads
Computer Engineering	245	117
Mechanical Engineering	756	147
Industrial Engineering	245	49
Electrical-Electronic Engineering	518	87
Civil Engineering	358	46
Total	2122	446

Table 4. 10: The number of ads that accept only one department

Table 4.11 shows the number of ads that accept applications from both one engineering department covered this research and one or more departments remaining outside of research:

Department	Turkish Ads	English Ads
Computer Engineering	88	102
Mechanical Engineering	301	55
Industrial Engineering	514	161
Electrical-Electronic Engineering	154	52
Civil Engineering	387	24
Total	1444	394

Table 4. 11: The number of ads that accept applications from multiple departments

4.1.1.5 Relative of job placement potantials for each engineering department:

We use this formula to estimate the potantial of job placements for each engineering department:

the total number of personnel / number of graduated students

Table 4.14 shows job placements potentials for each department.

- To estimate number of graduated students we use the 4-year-old student capacity of each engineering department in Turkish Universities.
- Software engineering is included in computer engineering numbers.
- Electrical engineering is included in electrical electronics engineering numbers.

The data of 2010 university placement data is used to create Table 4.12. We assume that all students settled in these engineering departments graduate at the end of 4 years.

The total number of student capacities in Turkish universities as of 2010. The data is from OSYM tables.	
Mechanical Engineering	9386
Electrical-Electronics Engineering	8129
Computer Engineering	7496
Civil Engineering	7102
Industrial Engineering	5207

Table 4. 12: The total number of student capacities in Turkish universities as of 2010.

Total Number of Personnel for Turkish & English Ads	
Mechanical Engineering	1721,3
Electrical-Electronic Engineering	1340,1
Industrial Engineering	992,4
Computer Engineering	896
Civil Engineering	830,7

Table 4. 13: Total Number of Personnel for Turkish & English Ads

Industrial Engineering	992,4/5207	0,190
Mechanical Engineering	1721,3/9386	0,183
Electrical-Electronics Engineering	1340,1/8129	0,165
Computer Engineering	896/7496	0,120
Civil Engineering	830,7/7102	0,117

Table 4. 14:Relative job placements potentials:

4.2 CONCLUSION

In this section we try to analyze ads which can only accept degree from engineering. We've classified 10.140 ads for most popular 5 engineering departments. We do the classification process automatically by using a code written in java looking for the keywords with pre-defined parameters in the ad texts. The classification results are shown in Table 4.3 and error rates are presented in Table 4.5. The total number of personnel for each engineering department is shown in Table 4.8 and Table 4.9. After the classification process, we calculate the job placement potential for these 5 engineering departments. For calculating the probability of job placements for each engineering, the total number of staff we found, compared with the total number of students that graduate from the universities in 2014. We assume the total number of students enrolling in universities in 2010 for this 5 engineering departments as the total number of students will graduate in 2014. So we obtain the result shown in Table 13.

As seen in Table 4.14 engineering departments were ranked from large to small based on their employment potential. According to the table, industrial engineering has the highest potential. The most important reason for this result is the number of graduated students from University for industrial engineering. It is less than the number of others. Just with this values we have obtained we can't reach definitive conclusions about the employment potential of the engineering departments. Because most of the ads look for experienced staff and most of the demanded skills will not be acquired at the university so employment potential may change with different data. We, however, rather than the result value on the table, we are interested in the ranking on the table. So, with the data that we obtained about the employment potential of the engineering departments although we may not reach definitive conclusions, we believe that on the table the ranking order is important.

CHAPTER 5

CLASSIFICATION OF ADS THAT ARE NOT EXPLICITLY SPECIFIED DEPARTMENT NAMES

5.1 Classification of ads that are not explicitly specified department names

In this chapter we classify ads that don't contain any university department names explicitly. These ads don't specify the type of personnel they are looking for. We predict the type of personnel they are looking for by examining the skill sets on the ads. In addition, we calculate the relative job placement potentials for departments including the data from this classification.

In this section, the number of ads we are going to classify for each department are shown in Table 5.1. We explained in chapter 3 how we obtained this data.

Mechanical Engineering	**1651**
Industrial Engineering	**2356**
Electrical-Electronics Engineering	**301**
Civil Engineering	**814**
Computer Engineering	**1319**
Total	**6441**

Table 5. 1: The total number of remaining ads

We prefer to classify ads based on engineering department separately. So we classify ads belonging to each department separately in the table. Some ads may have been retrieved for more than one department. In that case, those ads will be classified by more than one department.

5.1.1 Classification Method

We used Naïve Bayesian algorithm based on Chi-Square.

5.1.1.1 Naïve Bayes:

The Naïve Bayes is a simple probabilistic classifier based on applying Bayes theorem, and It has very simple implementation and a linear computational complexity. [11] It is

widely used for text classification. The basic assumption of Naïve Bayes is each word in the document is drawn independently from a distribution. [12],[13]

When the Naïve Bayes classifier is applied in Text Classifier problem we used this equation.[14]

$$p(class|document) = \frac{p(class).p(document|class)}{p(document)}$$

Where:

P(class|document): It's the probability of class given a document, or the probability that a given document D belongs to a given class C.

P(document): The probability of a document, we can notice that p(document) is a Constance divider to every calculation, so we can ignore it.

P(class): The probability of a class (or category), we can compute it from the number of documents in the category divided by documents number in all categories.

P(document|class) represents the probability of document given class, and documents can be modeled as sets of words, thus the p(document|class) can be written like:

$$p(document|class) = \prod p(wordi|class)$$

Where:

P(wordi|class) : The probability that the i-th word of a given document occurs in a document from class C and this can be computed as follows:

$$P(wordi|class) = (Tct + \lambda)/(Nc + \lambda V)$$

Where:

Tct: The number of times the word occurs in that category C.

Nc: The number of words in category C.

V: The size of the vocabulary table.

λ: The positive constant, usually 1, or 0.5 to avoid zero probability.

The performance of traditional Naïve Bayes is not as good as some other statistical learning methods such as nearest-neighbor classifiers [15] and support vector machines [16].But Naïve Bayes is very efficient and easy to implement compared to other learning methods. Thus, many research try to improve the performance of Naïve Bayes in text classification tasks.[11], [17], [18], [19], [20], [21] One of the most important problems in text classification tasks is that the feature space is very high dimensional which significantly reduces the classification performance.[22]. Besides the high dimensionality of the feature space, text classification problems are also characterized as frequently having a high degree of class imbalance.[23], [24]. To solve these problems in the study of literature selection feature

is usually performed as a pre-processing step. There are many research in literature to improve the text classification via focusing on feature selection problem.[23] [25].

There are many feature selecting algorithms. The most commonly used: Information gain, mutual information, chi-square, odds ratio, relevancy score. We used chi-square to select the features. Because chi-square is often used with Naïve Bayes to improve accuracy.[26], [11], [18], [19], [20]. In most of the empirical studies conducted for text classification such as [27], [28], [29], [30] and [31] chi-square is reported to perform better than many of its competitors.

5.1.1.2:Chi-Square:

Chi-Square is a very simple statistical technique used for feature selection. The Chi-square statistics tells us how relevant a word is to each class, and we will remove from the features, the words that are not relevant for that class. At the end, terms of high relevance are chosen. The formula for the Chi-square statistics is as follows in equation below [26]:

$$x^2(t,c) = \frac{N(AD - CB)^2}{(A + C)(B + D)(A + B)(C + D)}$$

This is the value for a term (t) and a category(c) where:

A: is the number of documents of category c containing the term t;

B: is the number of documents of other category (not c) containing t;

C: is the number of documents of category c not containing the term t;

D: is the number of documents of other category not containing t;

N: is the total number of documents.

5.1.1.3:Classification Process

We have classified the ads collected for each engineering department separately. We used binary classification. For example, while we make the classification for mechanical engineering, we only use ads in mechanical engineering folder. We classified these ads into two classes. If mechanical engineers can apply to this ad, we labeled it as "positive".

Otherwise, we labeled it as "negative". In this way, we classified all ads for 5 engineering departments. We classified English ads and Turkish ads separately.

We prepared training ad sets for "positive" and "negative" classes of each department. The positive training set contains the ads that accept applications from that department only with explicit department names. We prepared three negative training set. Two of them are for Turkish ads and one of them is for English ads. One of the negative training set contains the ads that accept applications from irrelevant departments with this study such as financial advisor, specialist, teacher etc. One of the negative training set contains some ads, which accept applications from high school and/or vocational high schools. The negative training set that for English ads contains the ads that accept applications from both irrelevant departments with this study and high school and vocational high schools.

In addition due to we use binary classification we also use positive training sets as negative training sets. For example while we are classifying computer engineering ads, we used all positive training sets as negative training set except computer engineering positive training set.

	For Turkish and English ads		
Positive training sets	**Computer Eng.**	**111**	**118**
	Mechanical Eng.	**128**	**147**
	Electrical&Electronics Eng.	**132**	**86**
	Civil Eng.	**126**	**47**
	Industrial Eng.	**121**	**50**
Negative training sets	**Negative training set1**	**176**	**124**
	Negative training set2	**160**	

Table 5. 2: Size of training sets:

5.1.1.4 Classification Result

The results of classification summarized in Table 5.3:

Department	Positively

	classified Ads
Industrial Engineering	**769**
Computer Engineering	**711**
Mechanical Engineering	**221**
Civil Engineering	**82**
Electrical-Electronic Engineering	**71**
Total	**1854**

Table 5. 3: The results of classification

We calculate classification errors separately for each engineering. For example, we classified computer engineering ads as "positive" or "negative". To determine the percentage of this classification error we checked the 50 ads classified as "positive" and we checked the 50 ads classified as "negative". In this way, we determined the percentage of error for computer engineering. We calculated the classification error rates for Turkish ads only.

Table 5.4 presents the number of ads we've checked to detect the errors of each department and the number of incorrectly classified ads:

Department	The number of ads that we've checked		The number of incorrectly classified ads	Percentage of incorrectly classified ads & Compensated Error percentages	
Mechanical Engineering	Positive	50	8	16	8
	Negative	50	4	8	
Electrical-Electronic Engineering	Positive	30	4	13,3	13,3
	Negative	30	0	0	
Industrial Engineering	Positive	50	7	14	4
	Negative	50	5	10	
Civil Engineering	Positive	30	4	13,3	10
	Negative	30	1	3,3	
Computer Engineering	Positive	50	3	6	4
	Negative	50	1	2	

Table 5. 4:The error rate

5.2 Re-Calculating the Relative Job Placement Potentials for Engineering Departments

In this section we don't calculate the number of personnel of classified ads. Because, although the system classified an ad for example as an ad of Industrial Engineering, also other department may apply for that ad. A department name does not take place in most of ad texts. Therefore, in this section, we included the classified ads in the calculation with a constant we thought it would be more accurate.

In this chapter while we calculate the number of personnel of classified ads, we multiplied the total number of ads by 0.66 for each department. In this way, every ad we have included in the calculation at a rate of 2/3. We thought this was the most appropriate rate. Table 5.5 provides the number of staff we found with this method.

Department	Number of Personnel
Industrial Engineering	**507,54**
Computer Engineering	**469,26**
Mechanical Engineering	**145,86**
Civil Engineering	**54,12**
Electrical-Electronic Engineering	**46,86**
Total	**1223,64**

Table 5. 5: The number of staff we found with this method.

We calculated the relative potential of job placements by adding the number of personnel we have found in this chapter to the number of personnel we found in chapter 3.

Department	Total Number of Personnel	Relative Potential of Job Placements	
Industrial Engineering	**1470,04**	***1470,04/5207***	**0,28232**
Mechanical Engineering	**1867,16**	***1867,16/9386***	**0,19893**
Computer Engineering	**1365,26**	***1365,26/7496***	**0,182132**
Electrical-Electronic Engineering	**1386,96**	***1386,96/8129***	**0,170619**
Civil Engineering	**884,82**	***884,82/7102***	**0,124587**
Total	**6974,24**		

Table 5. 6: The relative potential of job placements

5.3 Diploma based personnel search vs skill based personnel search

In chapter 4, we find the total number of ads for each engineering department. These ads require a diploma received from these departments. In this chapter, again we find total number of ads for each engineering department but these ads don't requires a degree in those engineering departments. Finally, for each engineering field, we calculate the percentage by the total number of ads compared to these two information separately. Figure 5.1 shows the results for each department.

Department	Diploma based personnel search	Skill based personnel search	Total
Mechanical Eng.	2337	221	2558
Electrical-Electronic Engineering	1846	71	1917
Industrial Engineering	1887	769	2656
Civil Engineering	1061	82	1143
Computer Engineering	1235	711	1946

Table 5. 7: Total Number of Ads

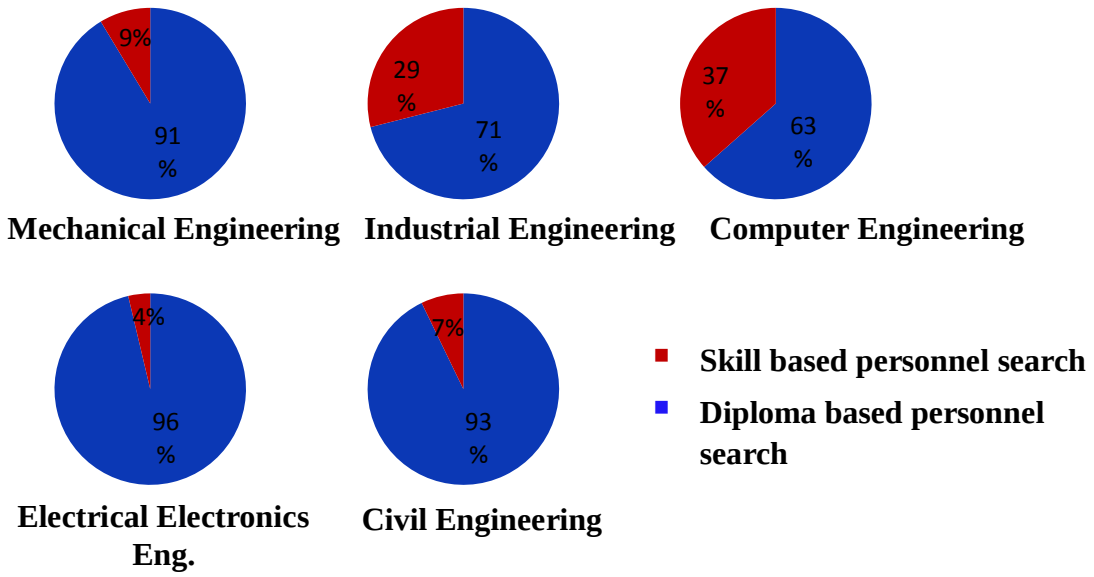

Figure 5. 1:Relative distribution of job ads

5.4 Finding popularity of engineering disciplines among university candidates:

In this chapter we conducted a popularity survey for 5 engineering departments according to minimum scores of the universities. We selected 10 state universities in Turkey. We used ranking of the entrepreneurial and innovative university index in 2014 for selecting 10 state universities. All of the selected universities have 5 engineering departments. Table 5.8 shows the selected universities and their minimum entry scores for 5 engineering departments separately.

Selected University	Computer Eng.	Electrical&Electronics Eng.	Industrial Eng.	Civil Eng.	Mechanical Eng.
Boğaziçi	519	531	526	498	515
Selçuk	338	367	334	367	317
Gazi	392	420	398	405	396
Erciyes	329	355	320	358	291
İstanbul	401	403	415	414	392
KTÜ	338	357	348	355	315
Atatürk	286	306	273	325	257

Pamukkale	316	347	332	350	305
19 Mayıs	324	343	327	355	310
ODTÜ	493	507	488	457	481

Table 5.8: Minimum entry scores of five engineering disciplines according to ÖSYM 2014 placement results

We ranked 5 engineering departments in each university's own. We obtain the result which shown in Table 5.9. When we sort 5 engineering departments for each university, we give points from 1 to 5 according to their degree in ranking. Then we add these points for each engineering department we obtain a popularity score as shown in Table 5.10.

Selected University	Computer Eng.	Electrical&Electronics Eng.	Industrial Eng.	Civil Eng.	Mechanical Eng.
Boğaziçi	3	5	4	1	2
Selçuk	4	5	3	5	2
Gazi	1	5	3	4	2
Erciyes	3	4	2	5	1
İstanbul	2	3	5	4	1
KTÜ	2	5	3	4	1
Atatürk	3	4	2	5	1
Pamukkale	2	4	3	5	1
19 Mayıs	2	4	3	5	1
ODTÜ	4	5	3	1	2
TOTAL	**26**	**44**	**31**	**39**	**14**

Table 5.9: Scores of Engineering Department

Department	Popularity Score
Electrical-Electronics Engineering	**4,4**
Civil Engineering	**3,9**
Industrial Engineering	**3,1**
Computer Engineering	**2,6**
Mechanical Engineering	**1,4**

Table 5.10: The Popularity Score

5.5 Conclusion

According to Table 3 Electrical and Electronics Engineering is the most popular department and Mechanical Engineering is the least popular department. According to this result we can say that there is no linear relationship between the popularity of departments and the job potential of departments. Because according to our findings, Mechanical Engineering has the second highest employment potential among 5 departments. And also we can say that most university candidates don't have any information about the employment potential of the department which to chose. We can say that there are some other factors affecting the choices of university candidates.

CHAPTER 6

ASSOCIATION ANALYSIS

6.1 ASSOCIATION ANALYSIS OF THE ENGINEERING DEPARTMENTS

In this chapter, we will evaluate the results of association analysis for the departments covered by this research. Our goal is to understand the relationship of an engineering department with the other engineering departments. To do this, we tried to find the number of ads which an engineering department occur together with the other engineering departments. We use FP Growth algorithm to find this. We explained this algorithm in Chapter 4.

Before the association analysis, we have implemented data pre-processing as discussed in Chapter 4. We eliminated rules to obtain the association rules that are important for us. The rules only takes place the name of departments together are important for us. To obtain all the association rules of departments covered by research only, we provide to be added only these departments' names to association tree. So, we have achieved to obtain the association that only included name of these departments.

We have assigned 1 to variable minSup and 5034 to variable maxSup. Due to the total number of files is 5034, we assign this value to maxSup because we try to obtain all association rules. In chapter 2, we have shown, as a table, the number of files that we have obtained for each engineering department as a result of the classification. When we combined these classified files in a single folder, we obtained 5034 unique files. So, these 5034 ads text contain engineering departments name explicitly. We use these 5034 ads to perform the analysis in this chapter.

As a result of the analysis, we determined the co-occurrence values in Table 6.1 for two departments. Figure 6.1 shows these vales graphically. These values primarily indicate the perspective of job advertisers. For example, if a job ad accepts applications from computer engineering and electrical-electronics engineering graduates. It means that the advertiser

considers this job to be handled by the graduates of both departments. In essence these values indicate the proximity of departments in the view of employers.

When we examined the results for the co-occurrence of two departments, we see that Computer engineering department is closest to electrical-electronics engineering department. Then it is closer to Industrial engineering department.

Electrical-electronics engineering department is closest to Computer Engineering and Mechanical Engineering departments with almost the same proximity. Industrial Engineering department is closest to Mechanical Engineering department.

One observation is that Civil Engineering is most isolated among these 5 departments. It occurs much less with these 4 departments. Other four departments are more related to each other.

Co-occurring Departments	Frequency
Computer Engineering, Electrical-Electronics Engineering	**368**
Computer Engineering, Industrial Engineering	**231**
Computer Engineering, Mechanical Engineering	**31**
Computer Engineering, Civil Engineering	**5**
Civil Engineering, Industrial Engineering	**45**
Civil Engineering, Mechanical Engineering	**139**
Industrial Engineering, Mechanical Engineering	**435**
Civil Engineering, Electrical-Electronics Engineering	**74**
Industrial Engineering, Electrical-Electronics Engineering	**193**
Mechanical Engineering, Electrical-Electronics Engineering	**370**

Table 6. 1:Frequency of Co-occurring Departments(2)

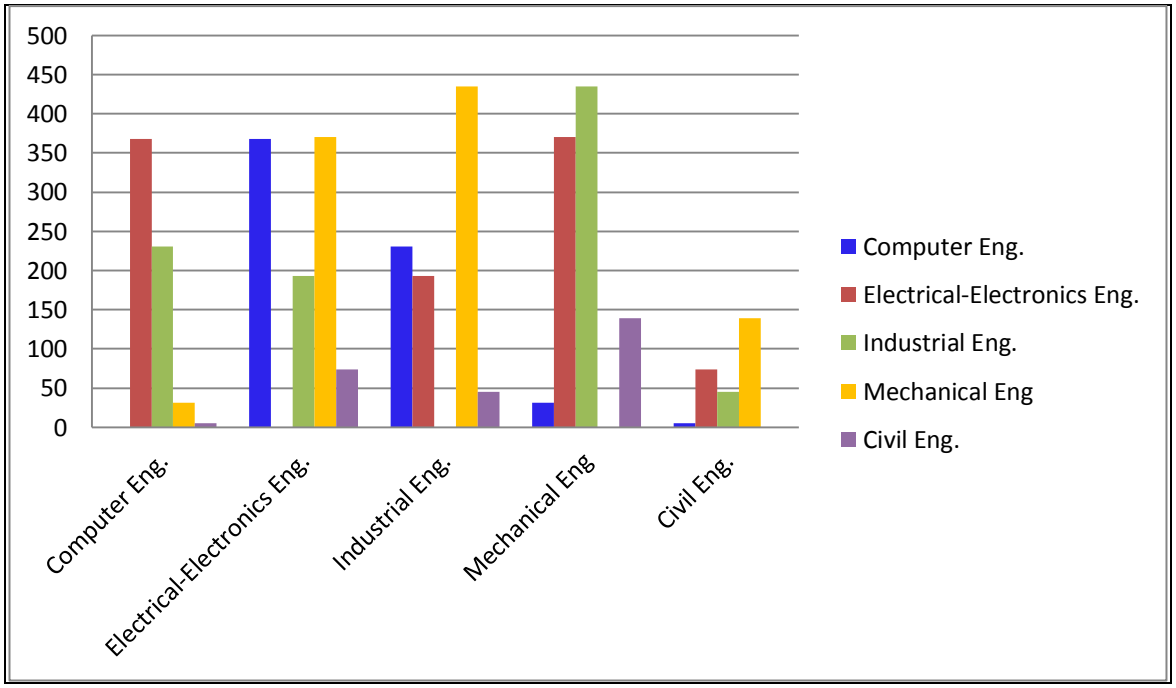

Figure 6. 1: Co-occurring Departments

Table 6.2 shows the co-occurrence values for three departments as groups. It has two values that is more than 50. The number of co-occurrence value is highest for Industrial-Mechanical-Electrical&Electronics triple. They occur 90 times together in ads. It shows that these three departments are closer to each other as a group of three.

Second highest occurring triple is Computer-Industrial- Electrical&Electronics. They occur 79 times together. It seems these three departments forms the next most relevant triple. Other triple groups have much fewer occurrences.

Co-occurring Departments	Frequency
Computer Engineering, Civil Engineering, Industrial Engineering	**1**
Computer Engineering, Civil Engineering, Mechanical Engineering	**1**
Computer Engineering, Civil Engineering, Electrical-Electronics Engineering	**2**

Computer Engineering, Industrial Engineering, Mechanical Engineering	**22**
Computer Engineering, Electrical-Electronics Engineering, Industrial Engineering:	**79**
Computer Engineering, Electrical-Electronics Engineering, Mechanical Engineering	**16**
Civil Engineering, Industrial Engineering, Mechanical Engineering	**29**
Civil Engineering, Electrical-Electronics Engineering, Industrial Engineering	**11**
Civil Engineering, Electrical-Electronics Engineering, Mechanical Engineering	**34**
Industrial Engineering, Mechanical Engineering, Electrical-Electronics Engineering	**90**

Table 6. 2: Frequency of Co-occurring Departments(3)

Table 6.3 shows the co-occurrence values for four departments as groups. There are two groups that are significantly higher than the others. Most frequent group is Computer-Industrial-Mechanical-Electrical&Electronics. It leaves Civil engineering outside. This is consistent with our conclusion that Civil Engineering is most isolated. The second most occurring group is Industrial-Mechanical-Electrical&Electronics-Civil. This leaves Computer Engineering outside. This result implies that Computer engineering is second most isolated department among these 6.

There is no ad that accepts applications from all these 5 departments.

Co-occurring Departments	Frequency
Computer Engineering, Civil Engineering, Industrial Engineering, Mechanical Engineering	**0**
Computer Engineering, Civil Engineering, Electrical-Electronics Engineering, Mechanical Engineering	**1**

Computer Engineering, Electrical-Electronics Engineering, Industrial Engineering, Mechanical Engineering	**13**
Civil Engineering, Electrical-Electronics Engineering, Industrial Engineering, Mechanical Engineering	**9**
Computer Engineering, Electrical-Electronics Engineering, Industrial Engineering, Civil Engineering	**0**
Civil Engineering, Computer Engineering, Electrical-Electronics Engineering, Industrial Engineering, Mechanical Engineering	**0**

Table 6. 3: Frequency of Co-occurring Departments(4)

6.1 ASSOCIATION ANALYSIS OF THE OTHER DEPARTMENTS

In the previous section we examined relations of 5 engineering departments with each other. In this section we will examine 5 engineering departments and their relationship with other departments. We perform association analysis process as explained above. We use FP Growth algorithm to find the number of ads which occur together one or more of 5 engineering departments which are involved by study with one or more of the university departments that are remained outside of the study. We perform the association analysis of each engineering department that we investigate, individually. For example, if we try to find out the associations of computer engineering with the other departments, while doing analysis we use computer engineering folders which we obtained in Chapter 4. And also we determined the university department names remaining outside the scope of research for each engineering department separately in chapter 4. So, if we try to find out the associations of computer engineering with the other departments, we use only university department names which we determined for computer engineering in Chapter 4.

Figure 6.2,6.3,6.4,6.5,6.6 shows the results for each engineering department. We assume math and mathematical engineering are the same department. We add the value of the two. And also we assume that management and management engineering are the same. All figures include the results we found in the previous section.

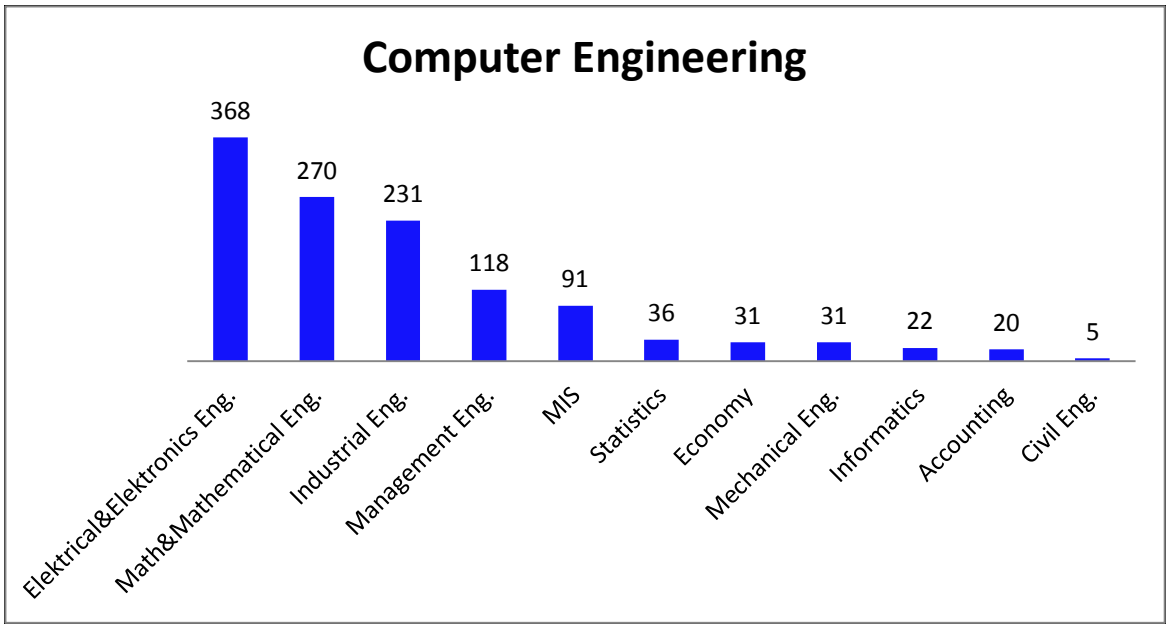

Figure 6. 2: The Result of Association Analysis of Computer Engineering

Figure 6.2 shows the number of co-occurrence of each department with computer engineering. When we examined the results for computer engineering we see that the sum of math and mathematical engineering is the highest value among the other departments This means that math is the closest department to computer engineering among the other departments remaining outside the study. But among all departments electrical&electronics engineering is the closest department to computer engineering.

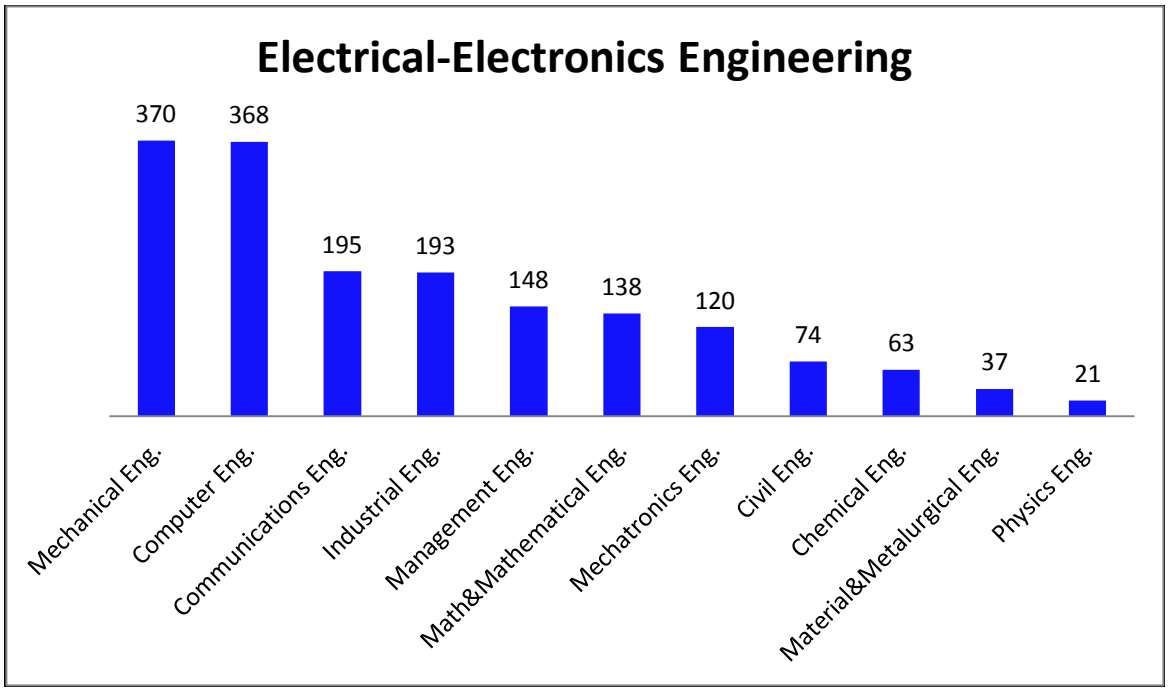

Figure 6. 3: The Result of Association Analysis of Electrical-Electronics Engineering

Figure 6.3 shows that still mechanical engineering and computer engineering are the closest department to electrical electronics engineering. Communications engineering is the third closest engineering department.

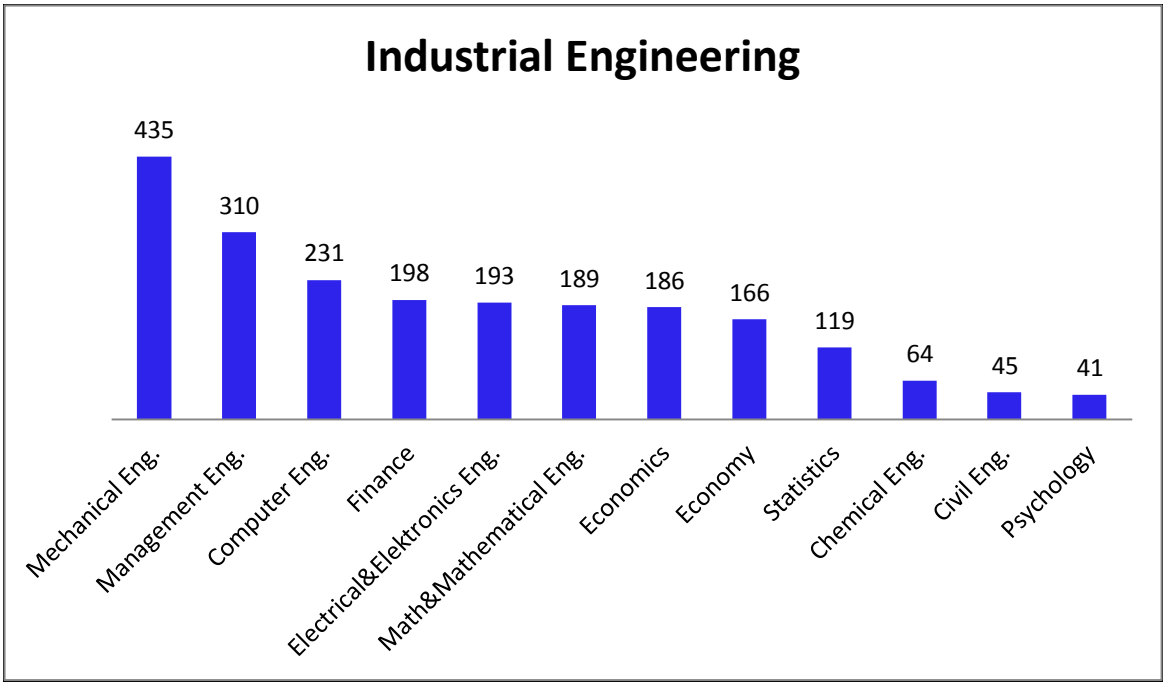

Figure 6. 4: The Result of Association Analysis of Industrial Engineering

Figure 6.4 shows the result for industrial engineering. As we can see easily from figure almost all values are over 100. This means that, compared to the other four engineering

disciplines, the number of closely relevant discipline for industrial engineering is more. Mechanical engineering and management engineering are the two closest engineering departments to industrial engineering.

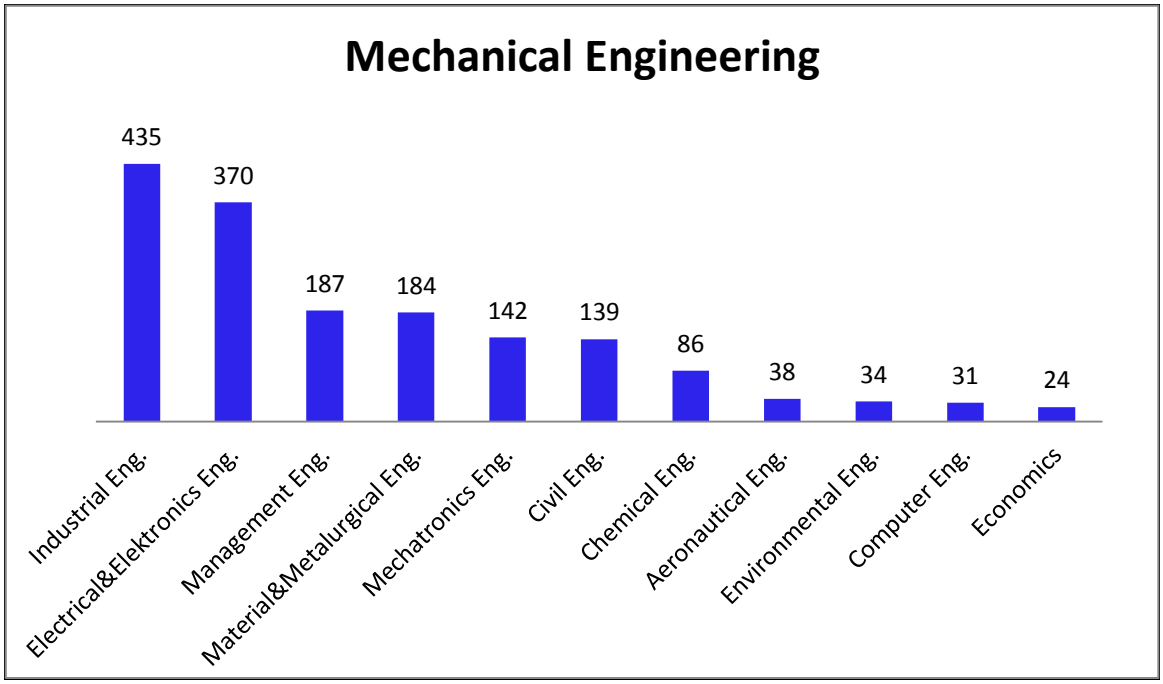

Figure 6. 5: The Result of Association Analysis of Mechanical Engineering

Figure 6.5 shows that industrial engineering and electrical electronics engineering are still most relevant departments with mechanical engineering and management engineering and materiel metallurgical engineering are the third closest department with almost the same value.

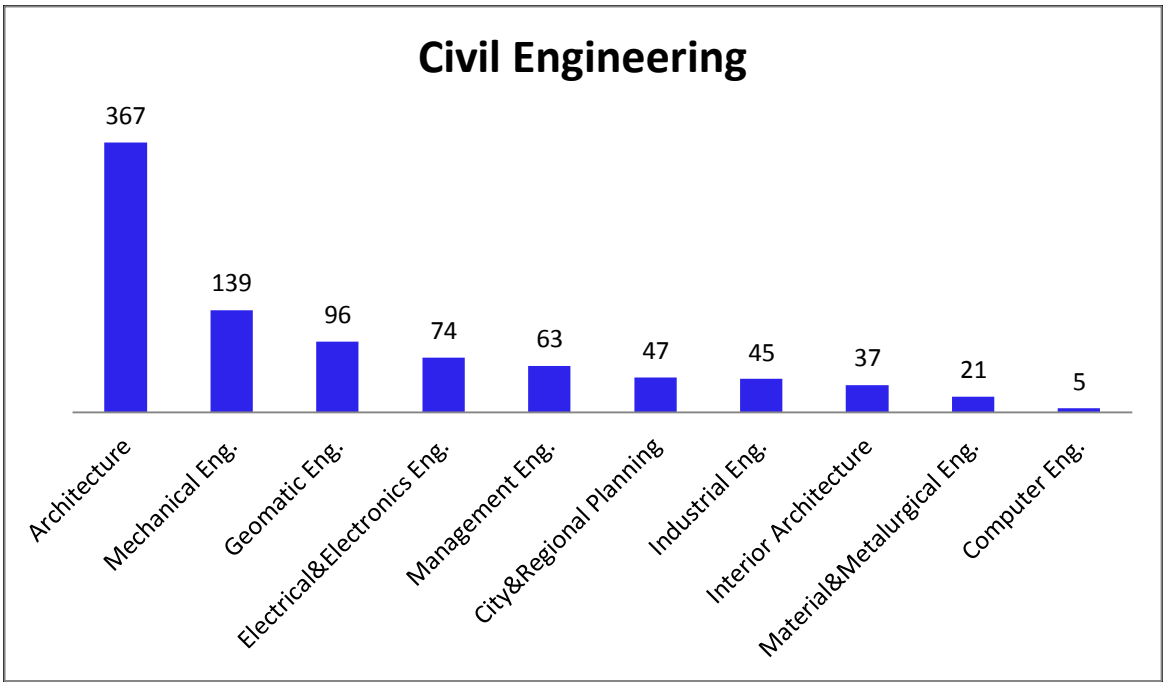

Figure 6. 6: The Result of Association Analysis of Civil Engineering

Figure 6.6 shows architecture is the most relevant department with civil engineering. Then it is closer to mechanical engineering department.

CHAPTER 7

FINDING THE MOST WANTED SKILLS FOR COMPUTER ENGINEERING

7.1 Finding the most wanted skills for Computer Engineering

In this chapter we define most wanted technical skills for computer engineering. For this we use computer engineering folders which we obtain in chapter 4 by collecting ads including computer engineering department names explicitly. We use FPGrowth algorithm for finding most occurred words in ads texts. As we discussed in chapter 4 we have prepared the data with pre-processing for using FPGrowth algorithm. But these pre-processes are not enough to finding skill terms. Because these pre-processes don't eliminate most of the unimportant words. Conversely, these pre-processes eliminate most of the important words. Because most of the technical skills are abbreviations and our elimination systems annihilate all abbreviations. This gives us an useful idea. So we decide to collect these abbreviations in a separate files. We find 294 abbreviations. But still some abbreviations are unimportant for us. We examine these abbreviations manually and decide which is a skill term. And we realize that some skill terms are not a Turkish word so we can distinguish these words via using some methods of zemberek. Thus, we obtain 443 words and examined all manually to decide which is a skill term. After long and tedious elimination processes finally 181 skill terms remained. Additionally we eliminate skills that appeared in fewer than 5.

Table 7.1 shows the skill terms we have found. And we have grouped these skill terms according to their similarities. Figure 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18 indicate the skill terms of each group and the number of ads of each skill term. Minimum frequency value for each skill terms is 5. Some skill terms which we cannot find any similarities among with the other groups, so we grouped this kind of skill terms as other.

Group Names	Technical Skill Terms
Operating Systems	Linux, Unix, Windows, Ubuntu, Solaris
Mobil Programming	Android, IOS, Objective-C, Interface Builder, Windows Phone
Frameworks	Entity Framework, Hibernate, JSF, Spring, Bootstrap, Struts
Database	SQL, ORACLE, PLSQL, TSQL, MYSQL, MSSQL, Stored Procedures, NoSQL, OLAP, ODI, ETL, ADO.NET, LİNQ, DB2, RDBMS, PostgreSQL, JDBC
Platforms	ASP.NET, J2EE, C#.NET, VB.NET
Programming Languages	JAVA, C#, ASP, PHP,C++, C, PYTHON, DELPHI, VB, XCODE, PERL, .NET, JAVASCRIPT
Client Side Scripting	JAVASCRIPT, JQUERY, AJAX, JSON
Web Server Side Programming	PHP, ASP, JSP
Network	IP, TCP, Firewall, WAN, CCNA, VPN, Active Directory, Switch, Cisco, DNS, NFC, Router, IPS, MCSE, Ethernet, CCNP, DHCP, Hosting RFC, SIEM, VoIP, SIP, HTTP, CEH, DLP, Nac, SSL,
Project Development Tools	SVN, TFS, GIT, Microsoft Sharepoint, CVS, Maven, IBM Rational, Jira
IDE	Eclipse, Netbeans, Visual Studio
Software Engineering	Agile, Scrum, Uml, SDLC
Server	Tomcat, IIS, Vmware, Apache, Weblogic, Load Balancing, WebSphere, Java Servlet,
Enterprise Resource Software	SAP, ERP, CRM, ABAP, Microsoft Dynamics, SD, PP, QM, FI, CO, HR, BW, BAPI, IDOC, ALV, AXAPTA, BADI, Netweaver, IFS, User-exit, Canias, SAP Basis, Netsis, PS, PI
Web Client Side Programming	HTML, CSS, HTML5, CSS3, XHTML, DHTML

Software Architecture	MVC, Design Patterns, CMMI, MVVM,SOA
Office	Office, Word, Excel, Access, Visio, Powerpoint, MSProject
Other:	SOAP, ORM, ITIL, Cobit, WCF, WPF, Rest, XML, Exchange Server, ARM, EMV, Crystal Reports, UX/UI Design, JPA, PMP, XSLT, Autocad, WSDL, ALE, IVR, Big Data, Oracle Discoverer, EJB, MCITP, Photoshop, Telerik, DevExpress, DSP, PCB, SSIS, Angularjs, CMS, EDI, Flash, PowerShell

Table 7. 1: Technical Skill Terms and Their Group Names

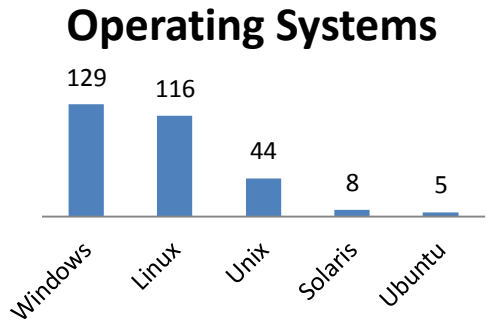

Figure 7. 1:Operating Systems

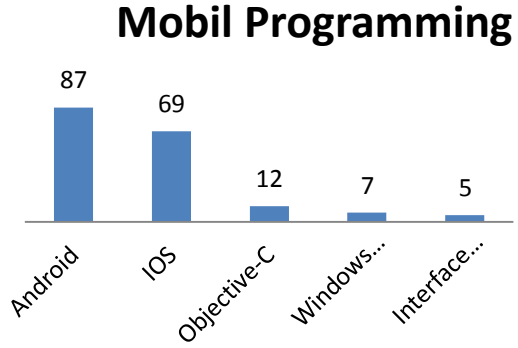

Figure 7. 2:Mobil Programming

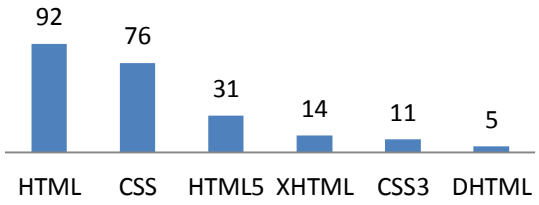

Figure 7. 3:Web Client Side Programming

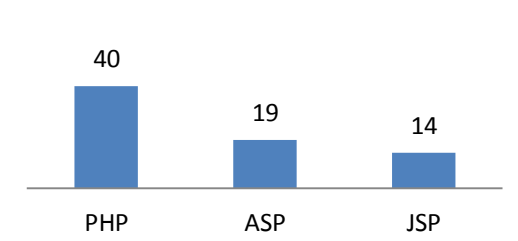

Figure 7. 5:Web Server Side Programming

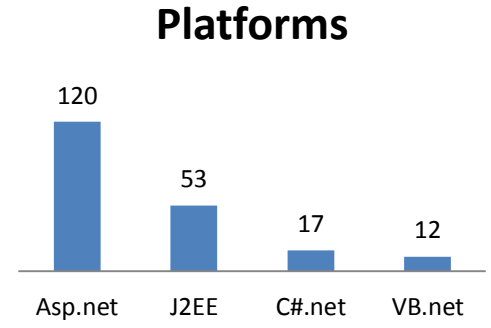

Figure 7. 4:Platforms

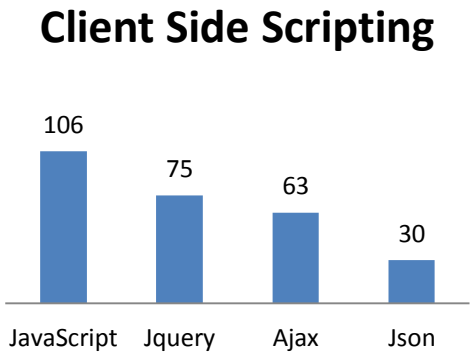

Figure 7. 6:Client Side Scripting

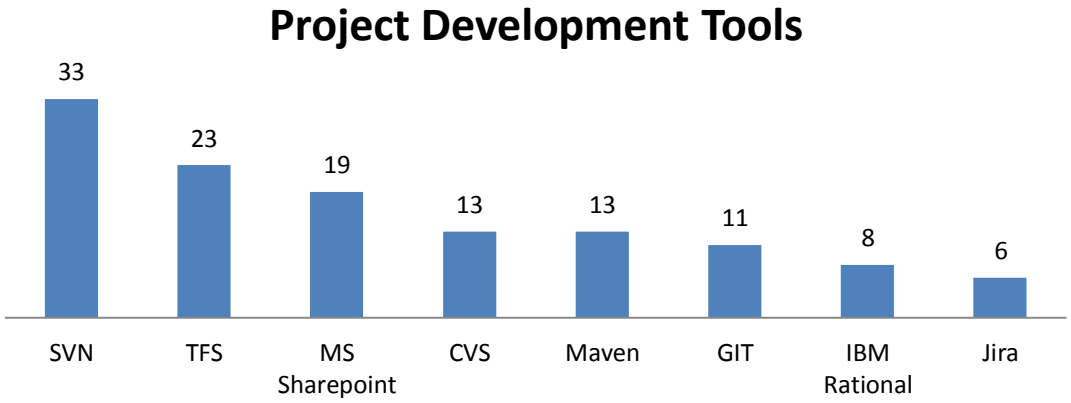

Figure 7. 7:Project Development Tools

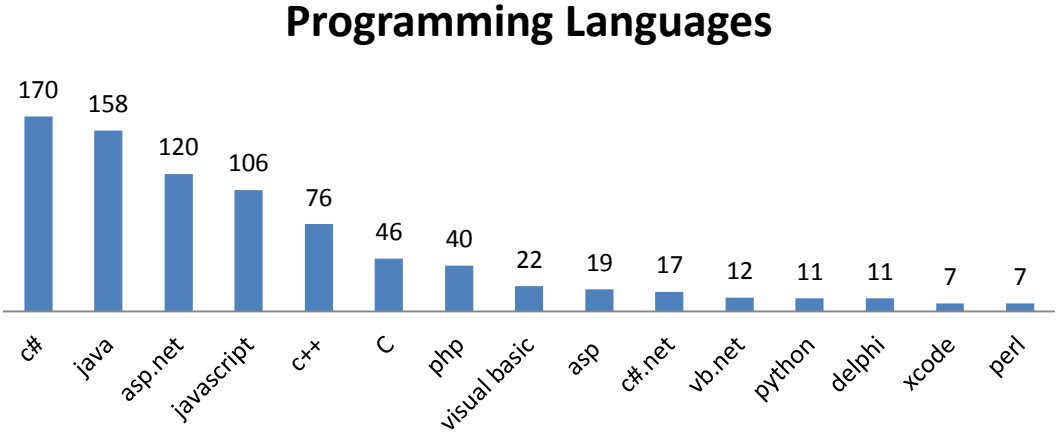

Figure 7. 8:Programming Languages

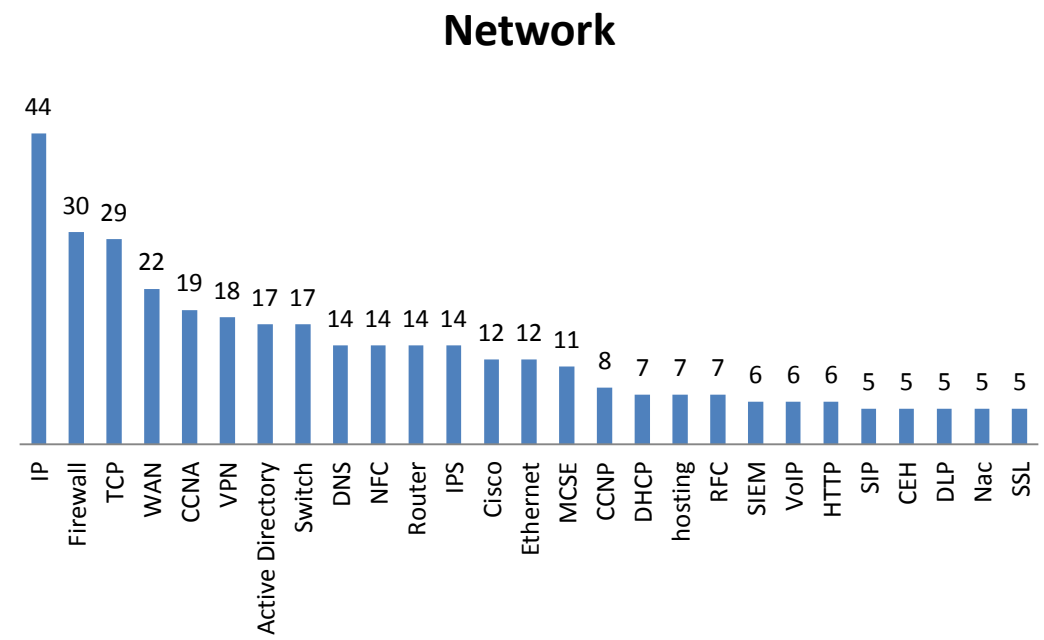

Figure 7. 9:Network

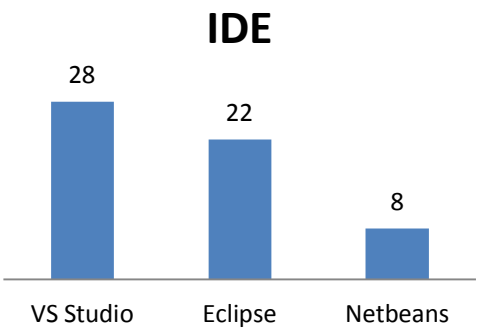

Figure 7. 10:IDE

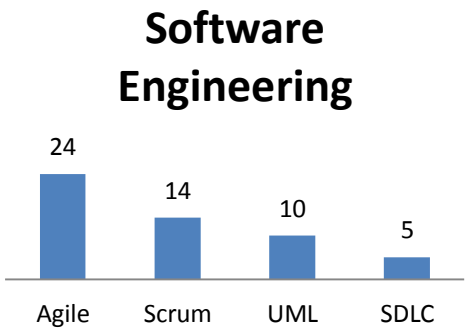

Figure 7. 11:Software Engineering

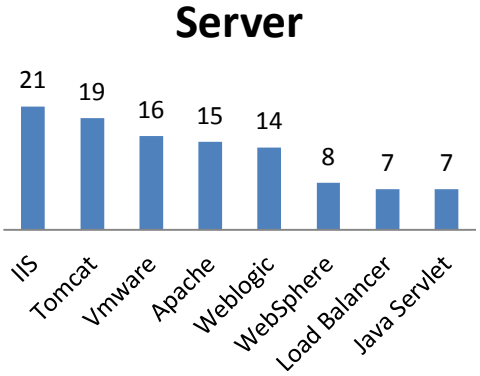

Figure 7. 12:Server

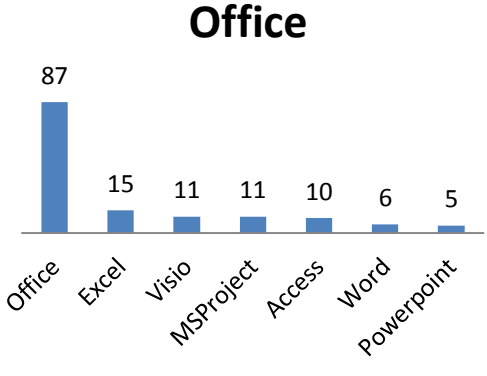

Figure 7. 13:Office

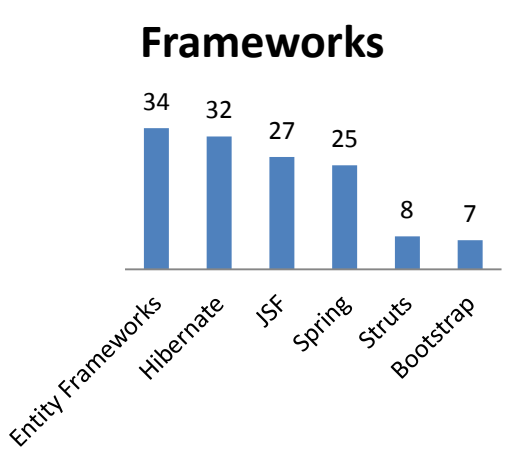

Figure 7. 14:Frameworks

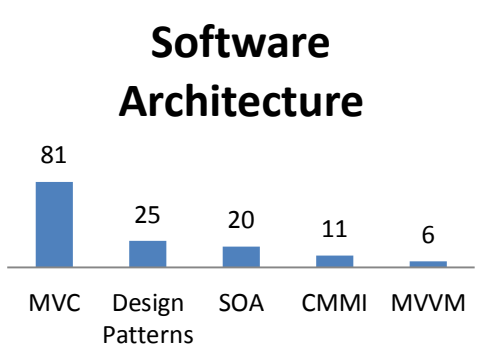

Figure 7. 15:Software Architecture

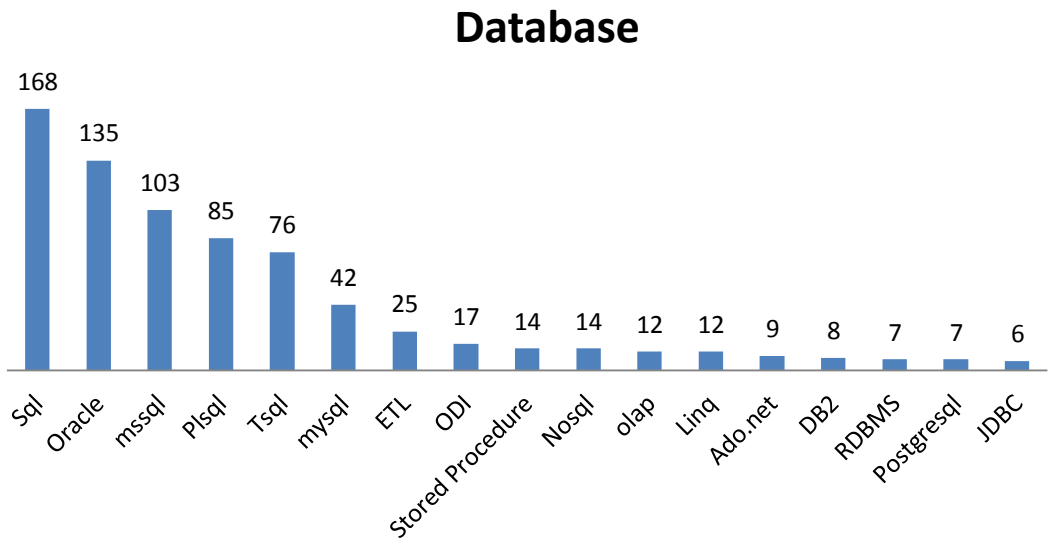

Figure 7. 16:Database

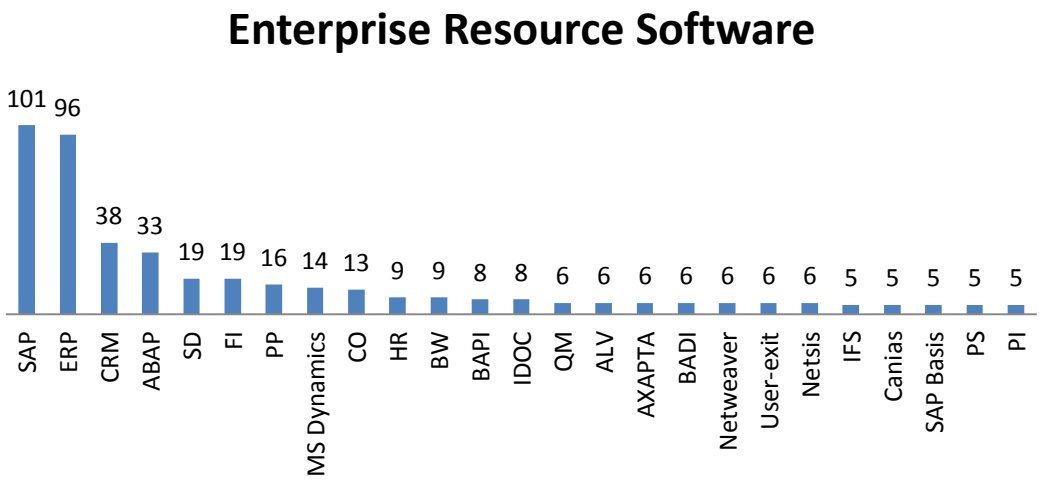

Figure 7. 17:Enterprise Resource Software

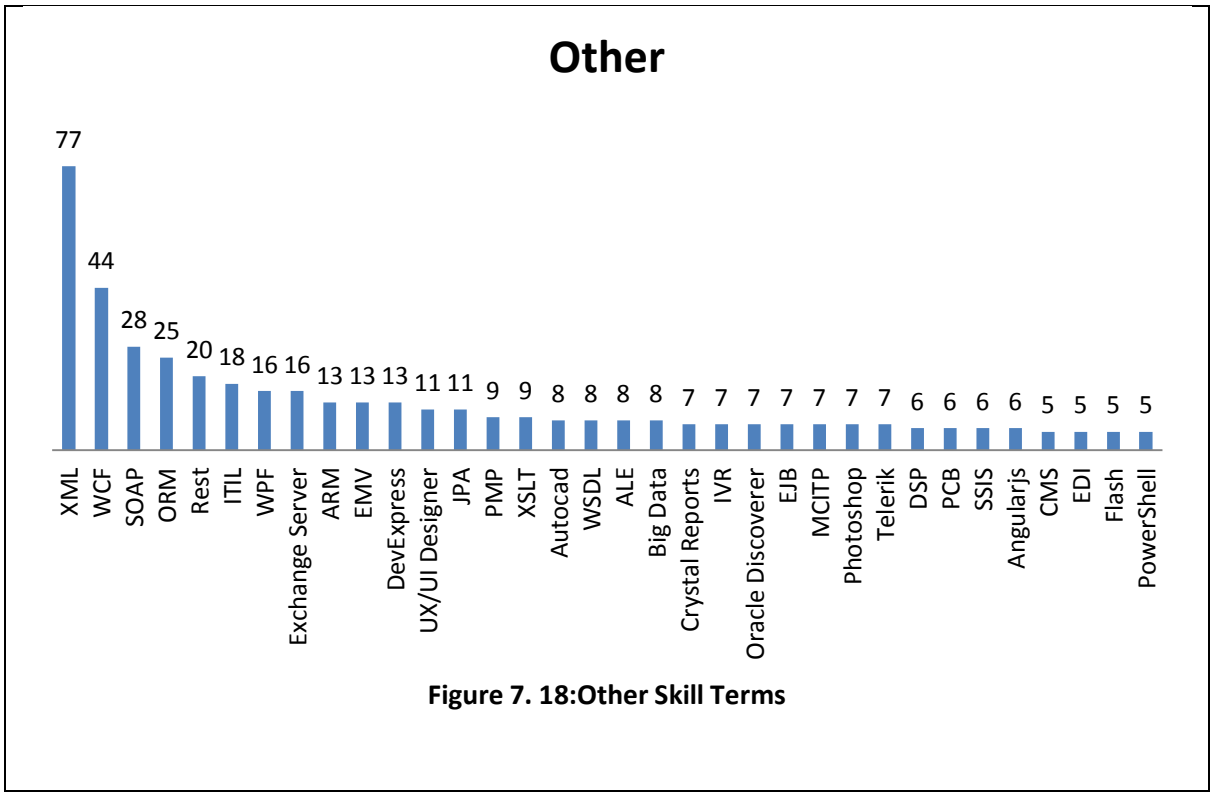

Figure 7. 18:Other Skill Terms

As we can see in Figure 7.1, although windows is the most important operating systems, if we add the frequency values of all other operating systems because all belong to the same family, the total value will be greater than the frequency value of windows.

Figure 7.2 shows Android and IOS are the most important programming skills for mobile apps.

Figure 7.8 shows C# and Java most wanted programming skills.

CHAPTER 8

CONCLUSION

2.1 CONCLUSION

In this study we examined the job ads using data mining techniques and have achieved significant results. Our aim is to reveal the statistical information hidden in the job ads. We think this information is important for many people in different areas. Although our study covers engineering departments also includes results for some other departments.

We calculated the relative job placement potentials for 5 engineering departments. The results are given at **Table 5. 8**. The list is as follows: Industrial Engineering, Mechanical Engineering, Computer Engineering, Electrical-Electronics Engineering and Civil Engineering. Industrial Engineering leads the list with a large margin. Then following three disciplines have similar values. At the end, Civil Engineering has the lowest job placement score.

In contrast, the popularity ranking of engineering departments among university candidates is quite different. It is given at **Table 5.10**. The list is as follows: Electrical-Electronics Engineering, Civil Engineering, Industrial Engineering, Computer Engineering and Mechanical Engineering. One might expect these two lists to overlap. However, the difference is quite significant. One reason for this big difference might be the lack of data to inform university candidates. They may have been selecting the university majors with insufficient data. Another reason might be that the job placement potentials is not a very important factor when selecting university majors. Though this should be unlikely.

Another result of this study presented at **Figure 5.2**. Diploma is more important for mechanical engineering, civil engineering and electrical-electronics engineering. However we can see the percentage of skill based personnel search is much more for computer engineering and industrial engineering. One reason for this, skills are at least as much as important diploma for this engineering departments.

We examined the relationships of the departments in job ads. Some of our main findings are: Computer engineering is the closest to Electrical-Electronics engineering. Civil engineering is the most isolated among these 5 departments. Because it is closest to Architecture. On the other hand, Industrial engineering co-occurs most with other departments.

Lastly, we conducted additional analyzes for computer engineering By this analysis we've found most sought-after talents for computer engineering.

2.1 LIMITATIONS

There are two limitations for this study. One of them, this study covers 5 engineering departments. And the other, this study represents the status of the job market in March-April 2015.

REFERENCES

[1] http://www.tuik.gov.tr/(Accessed on 30.11.2015)

[2] Huang, H., Kvasny, L., Joshi, K. D., Trauth, E. M., & Mahar, J.,"Synthesizing IT job skills identified in academic studies, practitioner publications and job ads." *Proceedings of the special interest group on management information system's 47th annual conference on Computer personnel research*. ACM, pp. 121-128, 2009.

[3] Smith, David, and Azad Ali. "Analyzing Computer Programming Job Trend Using Web Data Mining." *Issues in Informing Science and Information Technology* 11, pp. 203-214 , 2014.

[4] Smith, David, and Azad Ali. "ASSESSING MARKET DEMAND FOR WEB PROGRAMMING LANGUAGES/TECHNOLOGIES." *Issues in Information Systems , Volume 15, Issue II, pp. 411-420*, 2014.

[5] Litecky, Chuck, et al. "Mining for computing jobs." *Software, IEEE* 27.1, pp. 78-85, 2010

[6] Koong, Kai S., Lai C. Liu, and Xia Liu. "A study of the demand for information technology professionals in selected Internet job portals." *Journal of Information Systems Education* 13.1, pp. 21-28, 2002.

[7] Liu, Lai C., Kai S. Koong, and Les Rydl. "A study of information technology job skills." *37th Annual Southwest Region Decision Sciences Institute (SWDSI) Conference, Oklahoma City*, pp. 419-420, 2006.

[8] Webb, G. Kent. "The market for IS and MIS Skills and Knowledge: Analysis of On-Line Job Postings." *Issues in Information Systems, pp. 253-258*, 2006.

[9] Ayalew, Y., et al. "Computing knowledge and skills demand: A content analysis of job adverts in Botswana.", *International Journal of Advanced Computer Science and Applications,Vol. 2, No.1, January 2011*.

[10] R. Agrawal, T. Imielienski, and A. Swami. Mining Association Rules between Sets of Items in Large Databases. *Proc. Conf. on Management of Data*, 207–216. ACM Press, New York, NY, USA 1993

[11] Ting, S. L., W. H. Ip, and Albert HC Tsang. "Is Naïve Bayes a good classifier for document classification?." *International Journal of Software Engineering and Its Applications* 5.3 (2011): 37-46.

[12] Alpaydin E., Introduction to machine learning, 2nd ed. Cambridge, MIT Press, 2010.

[13] Aggarwal C. C. ve Zhai C., "A Survey of Text Classification Algorithms", Mining Text Data,Eds. Springer US, ss. 163–222, 2012.

[14] Thabtah, Fadi, et al. "Naïve Bayesian based on Chi Square to categorize Arabic data." *proceedings of The 11th International Business Information Management Association Conference (IBIMA) Conference on Innovation and Knowledge Management in Twin Track Economies, Cairo, Egypt*. 2009.

[15] Y. Yang and C.G. Chute, "An Example-Based Mapping Method for Text Categorization and Retrieval," ACM Trans. Information Systems, vol. 12, no. 3, pp. 252-277, 1994.

[16] T. Joachims, "Text Categorization with Support Vector Machines:Learning with Many Relevant Features," Proc. ECML-98, 10th European Conf. Machine Learning,

[17] Kim, Sang-Bum, et al. "Some effective techniques for Naïve Bayes text classification." *Knowledge and Data Engineering, IEEE Transactions on* 18.11 (2006): 1457-1466.

[18] Schneider, Karl-Michael. "Techniques for improving the performance of Naïve Bayes

for text classification." *Computational Linguistics and Intelligent Text Processing*. Springer Berlin Heidelberg, 2005. 682-693.

[19] Alexandrov, Mikhail, Alexander Gelbukh, and George Lozovoi. "Chi-square classifier for document categorization." *Computational Linguistics and Intelligent Text Processing*. Springer Berlin Heidelberg, 2001. 457-459.

[20] Meesad, Phayung, Pudsadee Boonrawd, and Vatinee Nuipian. "A chi-square-test for word importance differentiation in text classification." *Proceedings of International Conference on Information and Electronics Engineering*. 2011.

[21] Ikonomakis, M., S. Kotsiantis, and V. Tampakas. "Text classification using machine learning techniques." *WSEAS Transactions on Computers* 4.8 (2005): 966-974.

[22] Haltaş, Ahmet, Ahmet Alkan, and Mustafa Karabulut. "Metin Sınıflandırmada Sezgisel Arama Algoritmalarının Performans Analizi." *Gazi Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi* 30.3 (2015).

[23] Zheng Z., Wu X., ve Srihari R., "Feature Selection for Text Categorization on Imbalanced Data", SIGKDD Explor Newsl, Cilt 6, No. 1,80–89, Haziran 2004.

[24] Forman, George. "Feature selection for text classification." *Computational methods of feature selection* 1944355797 (2007).

[25] 27. Yang Y. ve Pedersen J. O., "A Comparative Study on Feature Selection in Text Categorization", Proceedings of the Fourteenth International Conference on Machine Learning, San Francisco, CA, USA, 412–420, 1997.

[26] Obasi, Chinedu Kingsley, and Chidiebere Ugwu. "Feature Selection And Vectorization In Legal Case Documents Using Chi-Square Statistical Analysis And Naïve Bayes Approaches." *IOSR Journal of Computer Engineering* 17.2 (2015): 42-50.

[27] 29.Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Machine Learning Research , 3*, 1289-1305.

[28] 30.Deng, Z.-H., Tang, S.-W., Yang, D.-Q., Li, M.-Y., & Xie, K.-Q. (2004). A comparative study on feature weight in text categorization. In *Advanced Web Technologies and Applications*, 588-597. Springer Berlin Heidelberg.

[29] 31.Debole, F., & Sebastiani, F. (2004). Supervised term weighting for automated text categorization. In S. Sirmakessis (Ed.), *Text Mining and its Applications*, 81-97. Springer Berlin Heidelberg.

[30] 32Man , L., Tan, C., Jian , S., & Yue , L. (2009). Supervised and traditional term weighting methods for automatic text categorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on , 31* (4), 721 - 735.

[31] 33.Man, L., Tan, C.-L., Low, H.-B., & Sung, S.-Y. (2005). A comprehensive comparative study on term weighting schemes for text categorization with support vector machines. *Proceeding WWW '05 Special interest tracks and posters of the 14th international conference on World Wide Web*, 1032-1033. New York, USA.