Creating Profile of Data Mining Specialist

Mirjana Pejić Bach
Mladen Varga
University of Zagreb, Graduate School of Economics & Business
E-mail: mpejic@inet.hr, mvarga@efzg.hr

Abstract. *The purpose of the paper is to present a survey on advertisements for data miner specialists published on the web site www.kdnuggets.com from May to December, 2004. The following characteristics have been investigated: industry, job title, education, years of experience, data mining tasks, data mining methods, software and additional skills. A typical data miner works in IT industry. His/her job title includes "manager", "data mining", "statistics" and "marketing" characteristics. An undergraduate or graduate diploma is required with less than 10 years of experience. The responsibilities include data preparation, building new models with SQL and SAS and presentation of the results. Classification methods are most often used. A data miner's most preferred additional skills are analytical skill, consulting, project management, customer profiling, and industry experience. The results of the survey have been used in profiling several courses at the Graduate School of Economics & Business of the University of Zagreb, Croatia.*

Keywords. Data mining, databases, course development

1. Introduction

Data mining (DM) is the process of finding new and potentially useful knowledge from data [www.kdnuggets.com] i.e. extracting or "mining" knowledge from large amount of data [Han, 2001].

A better term may be "knowledge mining from data" while mining refers to the content that is mined. There are a few other terms carrying similar or slightly different meanings, such as knowledge mining from database, knowledge extraction, data/pattern analysis, data archaeology, data dredging, but the most popular is knowledge discovery in databases (KDD).

The origins of data mining are various disciplines such as databases, statistics and artificial intelligence. The process of data mining consists of a number of tasks one person is not capable finishing alone.

The goal of the paper is to develop a profile of the data mining specialist based on the survey conducted from the job advertisements for experts in data mining published on the web site www.kdnuggets.com.

The results of the paper have been used in re-modelling the following courses: "Knowledge Discovery in Databases", "Analytical Data Processing", and "Business Intelligence" that are already offered or planned at the Graduate School of Economics & Business at the University of Zagreb, Croatia.

2. Data Mining

Data mining is the natural evolution of database technology that uses concepts, methods and techniques of various disciplines such as databases, statistics and artificial intelligence. The two major reasons that pushed DM as an important concept in information industry are:

- the availability of huge amount of data and the imminent need for turning data into useful information and knowledge
- the capability of information technology to cope with such amounts of data.

Today's information technology is capable of noting every activity in the form of digital data (for example after each telephone call a record is produced, any cash withdrawal is a transaction that leaves a record in a transaction database etc.). The amount of data collected in many databases is measured in gigabytes or terabytes.

The database industry has developed the following functionalities: *data collection*, *database creation*, *data management* (data storage and retrieval, transactional processing), and *data analysis* and *understanding* (data warehousing, data mining). The database technology has been evolving from primitive

file systems to sophisticated and powerful database systems that progressed from early hierarchical and network database systems to relational, extended relational, object-oriented, object-relational and deductive database systems that are used in the transactional information systems. In the decision support systems data warehouses as repository of multiple heterogeneous data sources organized under unified, mostly multidimensional, schema in order to facilitate decision making. Heterogeneous database systems and World Wide Web (WWW) as the global Internet-based information system has also emerged. Many application-based database systems, such as temporal, spatial, multimedia, active databases, knowledge bases and office information databases, play important roles.

The fast-growing huge amount of data, collected in large databases, has exceeded human ability for comprehension. Such databases have become data archives that are seldom used. The situation is characterized as data-rich but information-poor. The information is hidden in the stored data. Besides the information-rich data, the decisions are often made by intuition because of the lack of appropriate tools for extracting the valuable knowledge embedded in the vast amount of data.

Data mining helps to uncover important information and knowledge embedded in the data contributing greatly to decision making, business and science. The gap between data and information may be solved by development of data mining methods and tools.

3. Roles in Data Mining

Data mining is not a unique task as shown in Table 1, nor only one person does it. Data mining involves an integration of tasks from multiple perspectives: problem perspective, data perspective and method perspective, as shown in Fig. 1.

The problem perspective is important at the beginning and at the end of the data mining process. The person that has to be involved in these tasks may be just called the *user*. The tasks are: selecting the problem, defining the problem, evaluate the knowledge and apply the knowledge.

The data perspective involves all tasks in preparing data for the process of data mining. The *data provider* may be the title of the person performing these tasks. The tasks are: data selection, data pre-processing (cleaning, integration), and data transformation.

The method perspective comprises all tasks in the usage of the data mining methods, and the person carrying out these tasks could be called the *data miner*. The tasks are analyzing data, selecting the method, implementing the method, and finally presenting and interpreting the results.

Table 1. Tasks in DM

who (perspective or role)	what (task)
user	• select problem • (re)define problem
data provider	• prepare data
data miner	• select method • apply method • visualize results • interpret results =knowledge
user	• evaluate knowledge • apply knowledge

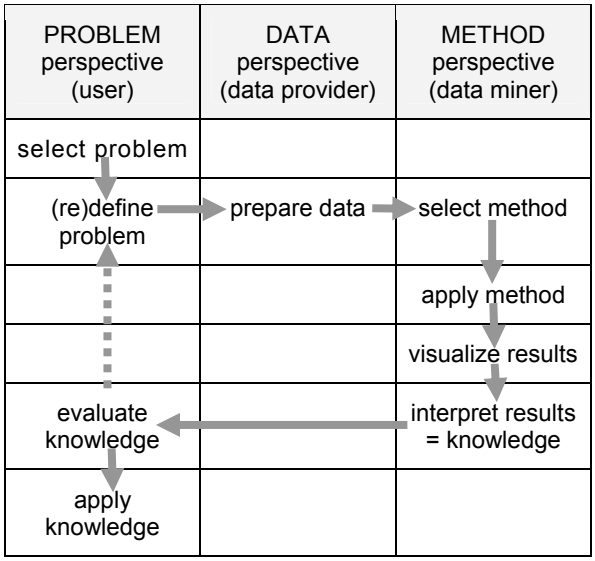

Figure 1. Perspectives (roles) in data mining

4. Survey of Advertisements for Data Miners

The web site www.kdnuggets.com is an well-known resource of data mining software, solutions, companies, web sites, publications, courses, and jobs. Firms are invited to find experts in data mining by advertising in KDnuggets. AOL, Amazon, eBay, Fidelity, Microsoft, SPSS, PriceWaterhouseCoopers,

T-Mobile, Verizon, and Yahoo are just a few of the leading companies that advertised their jobs there.

We have tracked advertisements during 8 months, from the beginning of May 2004 to the end of the year. During that period, 118 job advertisements were published. Table 2 shows industries searching for data miners. Most of the firms were in the information technology (IT) field (43%), marketing (19%), and banking and insurance (17%). Other industries comprise less than 10% of the total number of advertisements.

Table 2. Industry

Industry	#	%
IT	51	43%
Marketing	23	19%
Banking and insurance	20	17%
Healthcare	8	7%
Finance	6	5%
Trade	4	3%
Other	6	5%
Total	118	100%

It is not easy to recognize from the job title in the advertisement that the data miner is the searched specialist. Job titles are not unique (Table 3), so it was hard to classify them. In most cases job title contains the word "manager" (20%), while words "data mining" are mentioned in 16% of advertisements. There are about 10% advertisements that contain words "statistics", "operations research", "marketing analyst", and "software engineer/architect". Other job titles are mentioned in less than 5% of the advertisements.

The requirements for educational degrees and years of experience are heterogeneous. In 40% of advertisements an undergraduate degree is required, and the same percentage is true for the graduate degree (M.Sc. or MBA). A rather large percentage (20%) accounts for the Ph.D. requirement.

About a third of the companies searching for a data miner did not specify the required years of experience. The same percentage accounts for the companies searching for experts with less than 5 years of experience and 5 to 10 years of experience. More than 10 years of experience is required only in 3% of the advertisements.

Table 3. Position (job title)

Job title	#	%
Manager	24	20%
Data mining	19	16%
Statistics / Operations research	14	12%
Marketing Analyst	12	10%
Software engineer/architect	12	10%
Data warehouse/database developer	5	4%
Business intelligence analyst/developer	4	3%
Business analyst	4	3%
Consultant	4	3%
Research analyst	3	3%
Modelling analyst	2	2%
Data analyst/architect	2	2%
Other	13	11%
Total	118	100%

Although we may expect that a data miner would perform most of the data mining tasks, not all of them were required in the advertisements. A probable explanation is that most of these tasks are prerequisites for a data miner's job. As shown in Table 4, the presentation of the results is required in one third of the advertisements, building new models in 15%, and preparing data in 13%. About 10% of the advertisements described the jobs as carrying out the following tasks: selecting method, validating models, and improving the existing models. Less than 10% of the advertisements called for the skill of defining the problem. We conclude that the data miner is required to do all of these tasks, but effective presentation of the data mining results is especially desirable.

Table 4. Data mining tasks

Data mining task	#	%
Define problem	5	4%
Prepare data	15	13%
Select method	11	9%
Build new models	18	15%
Validate models	11	9%
Improve existing models	11	9%
Present the results	34	29%

The majority of the advertisements did not specify any data mining method. As shown in Table 5, 18% of the firms required classification methods, and we may conclude that the well-known statement that classification is still the bread and butter of data mining (Pyle, 2003) is true. The following

classification methods are required as well: logit regression, CART and CHAID decision trees, discriminate analysis, and neural networks. Prediction methods are specifically required by 8% of the advertisements, and these are: time series, linear regression, and neural networks. The same percentage goes to the undirected data mining methods (association algorithms, segmentation analysis, cluster analysis, and factor analysis), and statistical techniques/modelling.

Table 5. Data mining methods

Method	#	%
Classification methods	21	18%
Prediction methods	10	8%
Undirected data mining	10	8%
Statistical techniques/modelling	9	8%

More than 40% of firms that employ data miners require knowledge of SQL and SAS (Table 6). About 20% of the firms state in their advertisements that knowledge of C++, Unix and Excel is desirable. Oracle, Java, SPSS, Access and Perl are desired by 10-15% of the firms. Other software is mentioned in less than 10% of advertisements. It is interesting to note that specific data mining software, such as Clementine, is rarely mentioned.

Table 6. Software requirements

Software	#	%
SQL	52	44%
SAS	50	42%
C++	28	24%
Unix	22	19%
Excel	20	17%
Oracle	19	16%
Java	17	14%
SPSS	15	13%
Access	13	11%
Perl	12	10%
Software for web site development	10	8%
Other databases	9	8%
MicroStrategy	9	8%
Datawarehouse tools	9	8%
Clementine	6	5%
Visual Basic	4	3%
S-Plus	4	3%
Other software	32	27%

The firms searching for data miners also stress in their advertisements that additional skills are required. As shown in Table 7, the most often required are analytical skills (42%), consulting (19%), project management (16%), customer profiling (15%), and industry experience (15%). The experience in the following areas is mentioned: financial marketing, credit industry, CRM, market research, insurance industry, fraud detection, advertising, and database marketing.

Table 7. Skills required

Skills required	#	%
Analytical skills	50	42%
Consulting	23	19%
Project management	19	16%
Customer profiling	18	15%
Industry experience	18	15%
Direct marketing	15	13%
Database/data warehouse	14	12%
Management	13	11%
Machine learning	12	10%
Communication skills	11	9%
Working with clients	7	6%
Other	10	8%

5. Profiling Data Mining Courses

The results of the study were used to profile the existing undergraduate courses dealing with data mining, offered at the University of Zagreb's Graduate School of Economics & Business. These courses are: "Business Intelligence", "Analytical Data Processing" and "Knowledge Discovery in Data Bases".

Some questions, raised in profiling mentioned courses, were: What are the most important data mining tasks that have to be taught? What are the most useful data mining methods? What is the best software in teaching data mining? What skills are highly connected with data mining? What industry cases are useful to be explained in the courses?

Although all data mining tasks have to be presented in data mining courses, the study shows that the most important tasks to be considered thoroughly in the courses are preparing data, building data mining models, and presenting the final results.

The most important and the most used data mining methods are various classification methods. Among them decision trees/rules is the most popular (www.kdnuggets.com) and easiest to understand. Since it is not possible to present all data mining methods in the

courses we have chosen to present the decision trees/rules.

Good data miners must be skilled in two types of software. The first type is statistical software, such as SAS; and the second is the data manipulation software, such as SQL.

Some additional skills that have to be mastered by data miners and included in the courses are various analytical skills, user consulting and profiling skills as well as project management skills.

The greatest demand on the data miners has been identified in the fields of IT, marketing, banking, and insurance. Therefore, it is suitable to present the study cases from those fields.

6. Conclusion

The paper has reviewed advertisements for data miners published on the web site www.kdnuggets.com. This web site has very strong reputation among the data mining specialists, and we believe that the published advertisements objectively represent the current state of the field.

We examined 118 advertisements, tracked from May to December 2004. Most of the firms that want to employ data miners are in the IT industry (43%), marketing (19%), and banking and insurance (17%). Job titles under which data miners are employed are heterogeneous. However, the most frequent titles are: manager (24%), data mining (19%), statistics and/or operations research (14%), marketing analyst (12%), and software engineer/architect (12%). Equal number of firms want experts with undergraduate and graduate diplomas. Furthermore, a rather large percentage goes the Ph.D. degrees. Equal number of firms wants people with less than 5 years of experience, and people with 5-10 years of experience. Among typical data mining tasks the presentation of the results (34%), building new models (18%) and data preparation (15%) are most common. The knowledge of the classification methods is the most sought after (18%). However, prediction methods, undirected data mining methods and statistical techniques are equally required (8%). The most often required software are SQL (44%) and SAS (42%). The additional skills needed are: analytical skill (42%), consulting (19%), project management (16%), customer profiling (15%), and industry experience (15%).

Based on the results of the survey the profile of data miner is created. Data miners are the most probable employees in the IT industry (43%), the job title of many of them is manager (20%), their work is mainly connected with presenting the results and building data mining models (29% and 15% respectively), they most probably use classification methods (18%), and they are required to have analytical skills (42%). These results have been used in profiling several courses dealing with data mining at the Graduate School of Economics & Business, University of Zagreb.

References:

1. Han, J., Kamber, M. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, San Francisco, 2001.
2. Pyle, D. Business Modelling and Data Mining. Morgan Kaufmann Publishers, San Francisco, 2003.
3. www.kdnuggets.com [2005-02-10]