Chapter XIV
Data Mining of Personal Information:
A Taste of the Intrusion Legacy with a Sprinkling of Semantic Web

Dionysios Politis
Aristotle University of Thessaloniki, Greece

ABSTRACT

In this chapter data-mining techniques are presented that can be used to create data-profiles of individuals from anonymous data that can be found freely and abundantly in open environments, such as the Internet. Although such information takes in most cases the form of an approximation and not of a factual and solid representation of concrete personal data, nevertheless it takes advantage of the vast increase in the amount of data recorded by database management systems as well as by a number of archiving applications and repositories of multimedia files.

INTRODUCTION

It is a common secret that personal data that should be handled cautiously are "leaking" intentionally or unintentionally due to "mistakes", mismatches in Information Systems handshaking, or hacking. The last threat is the most difficult to cope with since the guardians in the inner circle have to supervise not only the hierarchical administrative structure that maintains and accesses the data, but also to watch carefully the Information and Communication Technologies advances that provide alternative routes to sensitive data for a variable number of support personnel.

However, technological innovations and multimedia gadgets of various forms have shaped another way to form personal data repositories. Indeed, in recent years there has been a vast increase in the amount of data recorded by database management systems as well as by a number of archiving applications. This explosion in the amount of electronically stored data was accel-

erated by the success of the relational model for storing data and the development and maturing of data retrieval and manipulation technologies. Apart from the large corporate databases which have been implemented, new forms of spreading information and storing data have emerged; the most notable of them are the World Wide Web (WWW) and the various multimedia databases formed out of diverse multimedia applications and presentations. The conjunction of hypertext languages and multimedia data such as images, video and audio, shape an immense network of information. It could be said that the WWW virtually is the largest database ever built.

While technology for storing the data has developed fast to keep up with the demand, little emphasis was paid to developing software for analyzing the data. It can be readily shown in Example 1 that the conventional Data Manipulating Language (DML) interfaces are insufficient or cumbersome in extracting statistical information.

Until recently, it was difficult not only to process but even to correlate information merging from diverse fields of data warehouses. The huge amounts of stored or tracked data contain knowledge that can be deduced, covering many aspects of the activities recorded as "raw" data. Database Management Systems in use at present manage these data sets allowing the user to access only information explicitly present in the databases.

EXAMPLE 1.
The following relation describes the students of a University with many departments according to the relational model. For each student a grade is recorded marking his overall performance.

```
STUDENT (SN(integer[6])PRIMARY KEY, NAME(string[20]), SEX (character),
MARK(integer[3]), DEPARTMENT(string[15], ADDRESS(string[40]), GRADUATION
(date))
```

Following this definition, sample tuple values for table `STUDENT` look like:

SN	NAME	SEX	MARK	DEPARTMENT	ADDRESS	GRADUATION
4567	BABOULI	F	74	LAW	THESSALONIKI	20-OCT-2003
7501	KINTS	M	85	RURAL ENGINEERING	VEROIA	15-JUN-2007
…	…	…	…	…	…	…

If we want a histogram on student performance deduced out of database records, we will have difficulty in **grouping by** student marks from 0 to 9, 10 to 19, 20 to 29, …, 90 to 100 using the standard Structured Query Language (SQL). Consequently, extensions to the SQL syntax have been proposed to simplify the task. For instance, the *N_tile* (Siberschatz et al, 1997) function supported by some database systems divides values into percentiles:

```
SELECT PERCENTILE, AVG(MARK)
FROM  STUDENT
GROUPBY N_TILE (MARK, 10) AS PERCENTILE;
```

N_TILE(`MARK, 10`) divides attribute MARK values into 10 consecutive ranges, with an equal number of values in each range; duplicates are not eliminated.