Computers and the Humanities **35:** 299–314, 2001.

Data Mining and Serial Documents

RACHID ANANE
School of Mathematical and Information Sciences, Coventry University
E-mail: r.anane@coventry.ac.uk

Abstract. This paper is concerned with the investigation of the relevance and suitability of the data mining approach to serial documents. Conceptually the paper is divided into three parts. The first part presents the salient features of data mining and its symbiotic relationship to data warehousing. In the second part of the paper, historical serial documents are introduced, and the Ottoman Tax Registers (*Defters*) are taken as a case study. Their conformance to the data mining approach is established in terms of structure, analysis and results. A high-level conceptual model for the *Defters* is also presented. The final part concludes with a brief consideration of the implication of data mining for historical research.

Key words: database, data mining, data warehouse, *Defters*, historical analysis, serial documents

1. Introduction

The perception that databases are a dormant potential resource is one of the driving forces behind the search for novel ways of analysing and processing large data sets in business, science and the arts. One of the new techniques for untapping resources in large databases is data mining. Its introduction owes much to the increasing reliance on database management systems and to the development of powerful computer architectures. Data mining is concerned with the uncovering and presentation of the underlying structure of the data source (Fayyad *et al.*, 1996a). Its considerable appeal is due to its ability to deal with different forms of data, such as spatial data, text-based data and multimedia data. Data mining is supported by a new generation of databases, called data warehouses, which are characterised by the integration of the data they contain.

This paper is concerned with the investigation of the relevance and suitability of data mining to historical serial documents in general, and Ottoman Tax Registers (*the Defters*) in particular (Singer, 1990). The uniform structure of the serial documents and the relative consistency of the data types they include make them ideal candidates for computerisation. The main thrust of this paper is that data mining is directly relevant to serial documents because the approach used in the analysis of these historical documents involves a number of steps that are conceptually similar to those applied in data mining.

The first part of the paper gives an introduction to the data mining process and establishes the need for data warehousing. In the second part, serial documents are introduced and their properties considered. The *Defters* are taken as a case study for the investigation of the suitability of serial documents to the data warehousing and data mining approach. This analysis is also supported by a presentation of a high-level conceptual schema for the *Defters*. A brief evaluation of the study concludes this paper.

2. Data Mining

Unlike traditional databases where the result of a query usually produces results that are either the extraction or aggregation from existing data, data mining is defined as:

> the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad *et al*., 1996b).

Data mining is basically concerned with the analysis of data using software techniques. It can use either a top-down verification-driven approach or a bottom-up discovery-driven approach (Simoudis, 1996).

A verification-driven approach does not create new information. It is an approach that is closely associated with traditional databases, and relies on query and reporting as the main operation, often in conjunction with simple statistical analysis. Its purpose is to validate a hypothesis expressed in terms of the entities and relations that exist in the database. This mode of enquiry is marked by the potential spawning of further queries in response to new insights.

A system based on the discovery-driven approach, on the other hand, is able to generate new concepts from existing information in the database. A discovery-driven approach can be used to predict or describe new information (Brachman *et al*., 1996).

Under prediction, the system is concerned with the identification of patterns in existing data in order to predict the future behaviour of some variable. This aim is supported by various techniques such as regression or classification methods.

A descriptive scheme requires greater user involvement for the interpretation of patterns, found and presented by the system. Description makes use of various techniques such as clustering and deviation detection.

These two approaches are supported by various tools and techniques such as decision trees, rule induction and neural networks. An application based on decision trees would use a hierarchy of *if-then* statements to classify data. It would segregate the data based on values of the variables. Rule induction, on the other hand, requires the generation of a non-hierarchical set of conditions which will then be used to predict values for new data items. This technique is heavily used in expert systems and can be used to create classification models. Unlike the two previous techniques, a neural network is a non-linear model that learns through training but has the disadvantage that the data mining process is opaque.

2.1. DATA WAREHOUSE

Whilst a database provides a framework for the storage, access and manipulation of raw data, a data warehouse is concerned with the quality of the data itself. Data quality is crucial to the search for patterns, and data mining draws its power from its symbiotic relationship with data warehousing:

> data mining can be done where there is no data warehouse, but the data warehouse greatly improves the chances of success (Inmon, 1996).

A data warehouse is defined as a

> subject-oriented, integrated, time-variant non-volatile collection of data (Inmon and Hackathorn, 1994).

Subject orientation refers to the fact that a data warehouse stores data about the important entities that are relevant to the business of an organisation. Data integration is achieved through a consistent encoding of key structures, naming conventions and the removal of duplicates. With integrated data the user can focus on analysis and data mining. Data is also time-variant because of its historical nature. It is indicated by the explicit association of time with the existing entities. Finally data is non-volatile because, unlike operational data in a database, it is never updated. The difference between a traditional database and a data warehouse is also marked by the structure of the data warehouse. A data warehouse includes the original detailed data from which the integrated data and the summarised data are derived. Summarised data is usually the result of previous research and enrichment to the database.

In addition to these different types of data, the data warehouse also incorporates metadata. Metadata is concerned with "the context of the information rather than the content". Its main function is to help the analyst navigate through the data warehouse and locate relevant information. It is a guide to the transformations required by the data warehousing environment and provides information about the relationship between the detailed data and the summarised data.

2.2. THE DATA MINING PROCESS

When the data warehousing phase is complete, data mining techniques can be applied to the data warehouse. The data mining activity is only one part of an iterative process that is made up of four steps (Simoudis, 1996):

1. data selection
   The first step is to determine the scope of the research and to select a subset of the data accordingly. If the historian is only interested in taxes, then only related data will be considered
2. data transformation
   Important transformations are suggested by the scope of the research and the data mining techniques used. Data may be organised in a specific way or enriched through the derivation of new values from existing attributes.

3. data mining
   The process of mining data involves the application of specific techniques to the transformed data. This includes verification-driven and discovery-driven operations
4. result interpretation
   The user is required to analyse and interpret the result of the data mining in the light of the scope of the enquiry. This step determines the need for further iteration.

In effect, the data mining process takes data from a data warehouse as input, applies data mining techniques to it and produces various results. The output of the data mining process can take several forms. These include facts, classification, trends, association or relation between two or more variables, patterns and deviation from the norm.

3. *Defters* and Serial Documents

In contrast to the complexity, irregularity and fuzziness of most historical documents (Denley, 1994), serial documents are composed of an array of comparable logical units with a common structure and purpose. Census records, in particular, provide a structure for regular data that lends itself easily to quantitative analysis. It is this property that highlights the importance and popularity of serial documents as subjects of study. According to one survey (Shürer *et al.*, 1992), the most frequently used source in Britain was the nineteenth century census:

> It reflects the ease with which a relatively simple data source may be transcribed into machine-readable form for subsequent analysis.

This affinity to modern models of storage and access favours the use of a model-oriented approach to database design. The semantic content of the textual and structural presentation of serial documents succeeds, to a large extent, in preserving the source (R. Acun *et al.*, 1994). This characteristic has led some historians to assert that:

> Documents which historians use, especially highly structured sources such as census returns, are in this sense already classified and coded databases (Higgs, 1990).

3.1. *DEFTERS*: PURPOSE AND STRUCTURE

The *Defters* constitute a major serial document and contain a considerable amount of data collected over several centuries for a large geographical area. Surveys for the *Defters* were conducted over a period of two centuries, from 1431 until the early part of the seventeenth century, at intervals varying from ten to thirty years. Detailed information about the population and the economic activity is expressed in terms of taxable revenue. Each province of the Ottoman empire was the subject of a separate survey and the results were recorded in a separate document. Existing archives include 1500 *Defters*, with each document containing an average of 400 pages.

The compilation of the *Defters* served two main aims. The first aim was primarily the identification of all sources of revenue, and the second was the distribution of the revenues to various beneficiaries. The mode of allocation of these revenues took three different forms: *mulk* (private property), *timar* (military fief) or *vakif* (religious institution).

A *Defter* is made up of two fundamental parts, the fiscal law (*Kanunname*) that governs a particular province and the listing of taxable revenue in that province. The typical categories of taxable revenue include a list of adults, information about communities, various totals from the settlement, a list of special agricultural land and breakdown of taxes, and the projected amount of revenue from each tax.

The *Defters* offer a relatively integrated collection of data with a consistent physical structure, adequate naming conventions for various categories of data, and "relative consistency across time-lines".

3.2. LIMITATIONS OF THE *DEFTERS* AS A 'DATABASE'

The *Defters* possess three important properties that are usually associated with a data warehouse. They deal with a specific subject, namely the taxable revenue. They are time-variant because they recorded, periodically, the surveys of a specific geographical area over two centuries, and their archival nature guarantees their non-volatility.

Yet, despite their relative consistency, they present two types of difficulty for the historian. The first type is inherent to the historical nature of the *Defters*. It includes the existence of context sensitive information, where a name may have different meanings in the same document. For example, the term *hane* could refer to a household, a married adult male or to a group of married and single males. Historical and geographical changes can also have a significant effect on location names. This difficulty is further compounded by the changes in surveying methods that led to the addition or removal of categories in successive surveys. It is a feature that is common to many serial documents, and Greenstein points to the inconsistency implied by these changes:

> In the UK, occupational data gathered by census takers changed dramatically during the 19th century while in the USA census takers continually refined and changed the categories used to take account of the foreign-born population (Greenstein, 1994).

The change in surveying practices was evident in the *Defters* (Singer, 1990). Although these changes are themselves historically significant they may, however, preclude the application of a universal procedure to the processing of serial documents. The task of the researcher is made even more challenging by missing pages in some *Defters*.

The second type of difficulty is closely linked to the purpose of the *Defters* as tax registers. The *Defters* deal mainly with aggregates when referring to production or population. Lump-sums were given without reference to a specific produce or individual units. Moreover, the surveys focused on a specific section of the population, the heads of the households who were usually male.

The first type of difficulty in the *Defters* points to the need for further integration of the data in order to ensure integrity and consistency across time-lines. The second type underlines the limitations of the *Defters* as a historical source and suggests the need for enrichment. This investigation helps identify three specific stages in the processing of the *Defters*:

1. data integration,
2. data enrichment and transformation, and
3. data analysis.

All these stages are supported by a specific historical context provided by the *Kanunname*.

3.3. *KANUNNAME*

The *Kanunname* performs three major functions in the *Defters*. It provides a map to the contents of the documents:

> The surveyors were also required to report on any local practice pertaining to that province (F. Acun, 1993),

acts as a guide to the tax calculations:

> It is indicated in the *Kanunname* of Rum that a double tithe was practised in the areas where a *malikane-divani* system was in force (F. Acun, 1993),

and may provide important information on population changes:

> Some evidence of the mobility of the population is provided in the law book (*Kanunname*) of the sub-district of Lidoriki in 1569. It was stated that: "as a result of their being tithed and harassed, the peasants were distressed and most of them were about to emigrate" (Doorn, 1989).

The *Kanunname* underlines the importance of context when analysing historical documents.

3.4. INTEGRATION

This important phase is concerned with the creation of an integrated and consistent data set with no duplicates. This requirement is particularly relevant to name changes:

> Because of spelling variations and place name changes, computerised toponymical research was carried out in order to systemise these changes and

> in order to link settlements in Ottoman registers both mutually and with later known villages and sites and former villages (Doorn, 1989).

The lack of consistency in terminology can also be a source of confusion and suggests the need for some abstraction when dealing with the *Defters*:

> Although some taxes bear different names, they are in fact concerned with the same object of taxation (Balta, 1989).

Missing values

The reconstruction of an accurate picture of the past implied in historical documents is often hampered by missing values. The *Defters* offer no exception to this rule. Unlike other documents, however, their serial nature and the stability of their structure allow for some reasonable extrapolation:

> In Granitsa, a village in the Lidoriki district, the number of families is incomplete in 1540, also due to missing pages. Because of its size in other years, the number of families was estimated to be 50 higher than the 14 mentioned in the source (corrected) (Doorn, 1989).

3.5. Enrichment

The focus of the *Defters* on taxable revenues points to valuable information that is left implicit in the documents. The need to have access to this untapped source is often expressed by the creation of "new context variables". The numerous studies of the *Defters* make use of two variables of interest: population and production. Although the documents give an indirect statement of these variables, their expression in terms of fiscal revenues enable the researcher to extract or aggregate new values from the *Defters*. This is achieved by the application of a number of operations to the data, such as enrichment and transformation which are often stated in terms of mathematical expressions. Thus, ratios, averages and coefficients correspond to enrichment, while various conversions are used to perform the transformations required by consistency constraints. The evaluation of the population and the production illustrates the use of these fundamental operations. They are indicated below in brackets.

Population

The population is expressed by fiscal units and does not include the non-taxpaying population of society, i.e. females, children and military people. In order to establish the size of the population, it is necessary therefore to use a coefficient to transform the fiscal units into a number of people. The household acquires a special status in the *Defters*:

> The number of people in the household was multiplied by various coefficients, ranging from 2.72 to 7 in order to obtain an approximate number for the family (F. Acun, 1993). [extrapolation]

This extrapolation provides the necessary means for determining the size of the population at large:

> On the basis of the number of settlements and families the average village was calculated (Doorn, 1989). [enrichment]

Production

Production as a focal point of study is, unfortunately, hampered by its implicit status in the *Defters*. From taxation and the mode of distribution of the fiscal revenues, in particular, it is possible to generate an explicit scale of ratios, thanks to the *Kanunname*. The enrichment process has, however, its limitations:

> Since fiscal revenues are expressed as aggregates, it is not possible to determine the production for each unit. As a result the estimates for each household can only be expressed as averages (Balta, 1989).

The determination of the production from the *Defters* in terms of meaningful units has posed a serious challenge to the historian. In his search for a solution, McGowan introduced a procedure for the evaluation of the production which laid the foundations for a number of quantitative studies (McGowan, 1969). The different stages of the calculations rely on the two fundamental operations identified above:

> I have added together all the tithe figures (monetary values) for grains, barley and millet and multiplied by 5 (since the rate of tithe is 1/5 in the area [*Kanunname*]) throughout the period under study to arrive at the total value for the production. [enrichment]
> Then I have divided this figure by the monetary value of one *mud* of wheat at the time concerned, i.e. by 60 for 1485 and by 120 for 1569 to arrive at a total agricultural produce in economic-wheat-equivalent (e.w.e). [transformation] from the total production I have deduced 33.33% for seed and milling losses. [enrichment]
> The result of this calculation was in *mud*, so I had to convert it to kilograms (1 *mud* is equal to 20 *kiles*, [transformation] 1 *kile* is equal to 25.656 Kg) [transformation]

3.6. *DEFTERS* AS A ‘DATA WAREHOUSE’

The enrichment of the data included in the *Defters* adds a new level of summarised data above the detailed data. Enrichment is often seen as a prerequisite for effective analysis:

> It is important to note that these transformations enlarge the range of existing data in the historical record but in a completely unambiguous way. Moreover, they provide a much more analytically powerful version of the data without losing the original. The advantage of the 'enrichment' of the database lies, of course in the creation of 'context' variables arising out of specific research interests (Collins, 1990).

The two pre-processing stages of the *Defters*, one concerned with data integration and the other with enrichment are essential to the transformation of the historical document from a mere subject-oriented, time-variant and non-volatile collection of data into one that is also integrated. The *Defters* would contain the original data, the integrated data and enriched data in addition to the fiscal law (*Kanunname*). As historical documents they display, therefore, a structure that is remarkably similar to that of a data warehouse.

3.7. *DEFTERS* ANALYSIS AND RESULTS

The *Defters* have been the subject of specific historical and geographical studies, where the spatial and temporal dimensions of the data have been used to great effect. The types of study range from the manual to computerised systems based on flat files or relational databases. The analysis process itself relies very often either on traditional statistical methods or on more sophisticated tools. Although historical analysis, as it is practised, includes aspects of the discovery-driven approach, such as classification, clustering or deviation detection, it is however mainly confined to a verification-driven approach.

Query refinement

Balta provides an interesting example of the historical investigation based on a top-down approach. The fundamental theme of her research is the determination of the net surplus in the village of L'Eubée. The refinement of this problem is subdivided into three sub-problems:

(a) to what extent did the production correspond to the needs of subsistence or taxation?

(b) to what extent did the economy of a village reflect the scope for communication between different regions or areas, and to what extent were these relationships achieved through money?

(c) to what extent was the net surplus destined to the fiscal needs and to what extent did they allow for net profits? (Balta, 1989).

The sub-problems can, in their turn, lead to a more concrete expression in terms of the entities found in the *Defters*. For a relational database, the refinement process will ultimately lead to the formulation of queries mapped onto SQL statements (R. Acun *et al.*, 1998).

Irrespective of the methods and techniques used in different modes of enquiry, the results and the findings of defterological studies are often represented by facts, classifications, associations, trends, patterns and deviation from the norm.

Facts

Facts constitute the primitive elements from which more sophisticated and elaborate structures can be built. Thus, Doorn found the simple fact that,

> In many villages, annual growth rates of over 2% were quite normal.

Facts can also be enriched by a spatial reference as follows:

> Also in Vitrinitsa the population grew, though not as fast as in Lidoriki,

or qualified by a temporal dimension:

> In Vitrinitsa the number of families doubled, whereas the growth of the Lidoriki district was less than 10% over the whole period from the 1520 till 1540 (Doorn, 1989).

Statements of facts in the *Defters* also highlight the importance of domain knowledge and historical context. Hütteroth found in the *Defters* that many villages in Anatolia were deserted after 1600. This fact was interpreted as a consequence of the Celabi riots which took place around 1600 (Hütteroth *et al.*, 1977).

Classification

Classification is an important outcome of a process of abstraction designed to endow a flat collection of data with a higher-level abstract structure. A classification is determined by a specific set of criteria. Suraiya Faroqhi makes her contribution to the study of the *Defters* by establishing the following classification for the towns of the sixteenth century Anatolia (Faroqui, 1979):

- towns with agricultural dues accounting for 40% of total revenue are agricultural market-towns,
- towns with commercial dues accounting for more than 40% of the total revenues are medium range or inter-regional market-centres,
- towns with commercial dues accounting for more than 75% are commercial towns.

Association

The power of association stems from its ability to bring into play more than one variable and to allow for a multi-dimensional analysis of the data. In the context of the *Defters*, the embedded link between population and production is further refined by the researchers, in terms of population growth on one hand, and commercial and agricultural activities on the other. According to Wagstaff (1985),

> By the middle of the 16th century, we see a considerable increase in commercial and agricultural activities. Manufacturing seems to have improved too. These developments, coupled with the increase in population, contributed to develop the village of Karahisar into a small size town.

The relation between agricultural activities and population growth is also confirmed by Cook in his study of 700 villages in Anatolia. The analysis of the ratio of arable land to population growth revealed that an increase in arable land was accompanied by population growth (Cook, 1972).

Trends

Trends possess an inherently temporal quality that is of special interest to the historian. Trends combine a dynamic description of the data with a powerful means for summarising, as shown by several studies of Anatolia. F. Acun, in her thesis on a region of Anatolia refers to commercial, manufacturing and agricultural activities in her search for meaningful trends in the sixteenth century. According to her results, in the middle of the sixteenth century, commercial activities represented 45% of the economic activity, manufacturing activities 15% and agricultural activities 38%. The second half of the sixteenth century is marked by an increase in the commercial and manufacturing sectors which hold respectively 47% and 19% of the economic activity. There is, however, a decline in agricultural activities (34%) (F. Acun, 1993).

Patterns

Patterns differ from other ways of presenting data by their richer semantic content. In the *Defters* they are often expressed by a combination of trend and association. At the heart of Faroqui's study of fifteen districts in Anatolia between 1520 and 1600 lies the search for patterns. The correlation between change in population and change in crop patterns was interpreted as population growth (Faroqui *et al.*, 1979). According to the authors, this population growth is manifested in the growth of commercialised crops and the reduction of wheat production. Unlike a mere static or fortuitous relationship between two variables, a pattern embodies a causal relationship:

> Is population pressure followed by a growth of commercialised crops?

Deviation from the norm

Deviation from the norm is often considered as the source of true discovery because it offers no immediate explanation (Fayyad, 1996b). Unexpected behaviour often points to further studies along new directions, as shown below by the case of the bachelors:

> In 1506 and 1521, however, the number of bachelors in Lidoriki was greatly reduced to only 3 or 4%. The decline was less dramatic in Vitrinitsa, where the

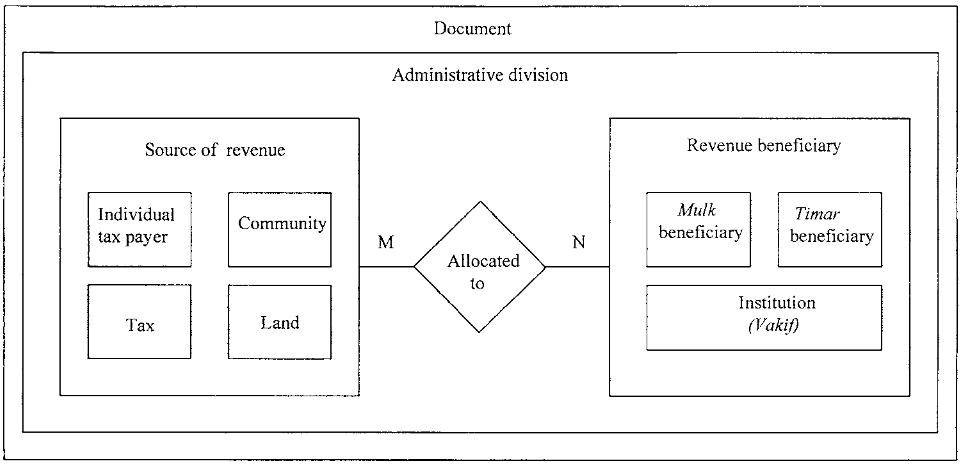

Figure 1. Purpose of the *Defters.*

rate in 1506 amounted to about 10%. Later the situation returned to normal in both districts (Doorn, 1989).

This interesting exception to a general pattern led to further research, and the author draws from sources outside the *Defters* in his search for an explanation:

> The extremely low proportion of married men was probably related to the *devshirme*, the recruitment of janissaries for the army. Passing through Lidoriki and Vitrinitsa on his campaign against Navpaktos in 1499, Bayezid must have recruited many janissaries (Doorn, 1989).

The importance of the search for patterns for the historian is that inconsistencies "point to new questions for study, which might otherwise not suggest themselves to the researcher" (Singer, 1990).

4. Modelling the *Defters*

Although the *Defters* contain a wealth of information that concerns a vast area of Europe and Asia, the scope of the application of computer methods to these documents has been restricted. Most of the models developed for the *Defters* have been the work of historians and are marked by an emphasis on numerical data. The main concern of the historians has been to organise and access the data in such a way as to allow the application of statistical methods. The underlying software models ranged from the specific, as in Door (1989) to a more flexible implementation such as that offered by Kleber (1990).

4.1. THE *TAHRIR* DATABASE

The *Tahrir* Database is a relational database which was developed for the fifteenth and sixteenth century *Defters*. The design and development of the *Tahrir* Database was motivated by the need for a general database for the *Defters*. Two constraints presided over its design. The first constraint was integrity of the source, and the second was the provision of software tools to support data mining. The conformance to these requirements owes much to the adoption of an abstract approach in the design. In Figure 1, a high-level conceptual model of the *Defters* identifies the main components of the documents. Allocation of resources is seen as the fundamental purpose of the *Defters*. It is represented by the relationship between two groups of entities: sources of revenue and their beneficiaries. A detailed presentation of this model and of the database is given in (R. Acun *et al.*, 1994).

This relational database was used extensively by F. Acun in her research on Anatolia in the sixteenth century (F. Acun, 1993). The approach was essentially verification-driven and relied heavily on the design and application of SQL queries. The study was a successful example of data mining on the original database using SQL only (Inmon, 1996).

Reliance on traditional data mining tools fails to take advantage of the temporal and spatial dimensions of serial documents. In the *Tahrir* Database, a significant effort was put into the development of software tools to support the interaction of the historian with the database. The provision of tools to support temporal and spatial analysis, in particular, is a step towards discovery-driven data mining. The *Tahrir* Database was enhanced by the development and introduction of a query system called HiSQL which extends the functionality of SQL in manipulating historical data (R. Acun *et al.*, 1998).

4.2. DATABASE ENHANCEMENT AND HISTORICAL RESEARCH

One of the central tenets of historical research is the preservation of the source. A software model should be as faithful as possible to the original document. It is evident, however, that the fuzziness, the incompleteness and the ambiguity of historical documents may be a source of incompability with software models. Relational database models, for example, require values for primary keys. This tension is further compounded by the fact that the documents deal with specific and concrete instances, whereas software models are informed by the principle of abstraction. In the *Defters*, for instance, different names were used for the same tax. Some annotation in the database was deemed necessary as a form of integration.

Integration, in this respect, plays a fundamental role in the preparation of the data for analysis, irrespective of the mode of processing. In software development, in particular, it can facilitate the mapping of historical documents onto the software models promoted by databases. Although this concession to software modelling may require some form of interpretation (and therefore may introduce bias), it does not compromise fundamentally the integrity of the source.

Enrichment and transformation, on the other hand, go beyond a mere conformance to software requirements or processing needs. The enrichment process creates a higher additional layer of information on top of the original database. Enrichment is, to a large extent, ‘interpretation-soaked’ and reflects the research interests of the historian. It can be realised by purely algorithmic methods or by the application of more elaborate queries on the database. The new information can be stored by generating new tables, thus effectively creating a new database.

In the case of the *Tahrir* Database, the implementation manages to preserve the original document. Support for data mining is provided by various software tools for searching and displaying information. Enrichment and further integration of the data can be performed by using these tools.

5. Serial Documents in Context

This investigation has shown that data preparation is the most important phase of the processing of the *Defters*. Whilst integration is considered critical to any analysis, irrespective of the means for processing data, historians are well aware of the tension and distortion that may result from enrichment:

> In this research we prefer not to convert the number of families into numbers of inhabitants, in order to keep the data as pure as possible (Doorn, 1989).

Within these historical constraints, integration and enrichment can produce a data set that conforms to the concept of a data warehouse. In this respect, the role of historical context provided by the *Kanunname*, as metadata, cannot be overemphasised. In addition, it was shown that the formulation of the results of several studies of the *Defters* can be mapped onto the output space of the data mining process.

Although the bulk of the studies under consideration focus on the results and ignore the tools of analysis, it is often the case that historians rely mainly on statistical techniques. In addition to their availability in packages, statistical techniques have also the advantage that they can accommodate both verification-driven and discovery-driven approaches. In his analysis of Medieval Farming Systems, Ken Bartley weighs the advantages and disadvantages of discriminant and cluster analysis, two relatively sophisticated techniques that represent two poles of data mining (Barley, 1996).

This example is indicative of the know-how of many historians and highlights the overlap that exists between data mining techniques and traditional methods. The set of techniques that are available to historians forms a subset, albeit small, of the large and sophisticated set of techniques used by the data miner, and thus ensures conformance on the analysis level.

This last point reinforces, therefore, the view that data mining and its symbiotic relationship to data warehousing are directly relevant to serial documents. The application of true data mining techniques to serial documents will hopefully provide a systematic framework for a more disciplined approach to the analysis of historical archives. Although this study has focused on the *Defters* the results are applicable to other serial documents.

The fundamental difference between the work of the historians and the modern data miner lies mainly in the use and the degree of integration of software technology.

6. Conclusion

The existence of historical archives and their increasing availability in computer storage media offer potentially exciting prospects for the application of data mining. The use and adoption of these sophisticated techniques is not without a price and may compromise the integrity of the source. They would require some awareness of their complexity in order to benefit fully from their potential. Historians should also be aware of the bias of data mining towards quantification and the need to resort to descriptive sources outside those being mined in order to obtain a fuller picture of an historical event.

The main conclusion, however, is that historians are in a similar predicament as the *Bourgeois Gentilhomme* of Molière: they have been doing data mining for a long time without being aware of it.

Acknowledgements

My thanks to Antoni Diller and Susan Laflin for their helpful comments.

References

Acun, F. *Ottoman Administration in the Sancak of Karahisari Sarki (1485–1569): An Analysis Based on the* Tahrir *Defters*. PhD Thesis, The University of Birmingham, 1993.

Acun, R., R. Anane and S. Laflin. “Database Design for Ottoman Tax Registers”. In *Yesterday*. Eds. H.J. Marker and K. Pagh, Odense, 1994, pp. 109–122.

Acun, R., R. Anane and S. Laflin. “HiSQL: A Front-end System for Historical Relational Databases”. *Computers and the Humanities*, 31 (1998), 483–502.

Balta, E. *L’Eubée à la fin du XVe Siècle. Economie et Population. Les Registres de l’année 1474*. Athens, 1989, p. 2.

Barley, K. “Classifying the Past: Discriminant Analysis and its Applications to Medieval Farming Systems”. *History and Computing*, 8(1) (1996), 1–10.

Brachman, R. *et al*. “Mining the Business Databases”. *CACM* 39(11) (1996), 42–48.

Collins, B. *Census Studies, Comparatively Speaking*. In *History and Computing III*. Eds. E. Mawdsley *et al*., Manchester University Press, 1990, pp. 184–187.

Cook, M.A. *Population Pressure in Anatolia, 1450–1600*. London, 1972.

Denley, P. “Models Sources and Users: Historical Database Design in the 1990s”. *History and Computing*, 6(1) (1994), 93–43.

Denley, P., Fogelvik S. and C. Harvery. *History and Computing II*. Manchester University Press, 1989.

Doorn, P.K. “Population and Settlements in Central Greece: Computer Analysis of Ottoman Registers of the Fifteenth and Sixteenth Centuries”. In *History and Computing II*. Eds. P. Denley *et al.*, Manchester University Press, 1989, pp. 193–208.

Fayyad, U. and R. Uthurusamy. “Data Mining and Knowledge Discovery in Databases”. *CACM* 39(11) (1996a), 24–26.

Fayyad, U., G. Piatetsky-Shapiro and P. Smyth. “From Data Mining to Knowledge Discovery: An Overview”. In *Advances in Knowledge Discovery and Data Mining*. Eds. U. Fayyad *et al.*, Cambridge, MA: MIT Press, 1996b, pp. 1–36.

Fayyad, U., G. Piatetsky-Shapiro and P. Smyth, R. Uthurasamy (Eds.). *Advances in Knowledge Discovery and Data Mining*. Cambridge, MA: MIT Press, 1996c.

Faroqhi, S. “Taxation and Urban Activities in the 16th Century Anatolia”. *International Journal of Turkish Studies* 1(1) (1979–80), 19–53.

Faroqhi, S. and Islamoglu-Inan. “Crop Patterns and Agricultural trends in Sixteenth-Century Anatolia”. *Review*, 2 (1979), 401–436.

Greenstein, D.I. *A Historian’s Guide to Computing*. Oxford University Press, 1994.

Higgs, E. *Structuring the Past: The Occupational and Household Classification of Nineteenth-Century Census Data*. In *History and Computing III*. Eds. E. Mawdsley *et al.*, Manchester University Press, 1990, pp. 67–73.

Hütteroth, W-H, and K. Abdelfettah. *Historical Geography of Palestine, Transjordan and Southern Syria in the Late 16th Century*. Erlangen, 1977, pp. 55–56.

Inmon, W.H. “The Data Warehouse and Data Mining”, *CACM* 39(11), 1996, 49–50.

Inmon, W.H. and R.D Hackathorn. *Using the Data Warehouse*. John Wiley, 1994, 2.

Kleber, A., “Computer-Aided Processing of Ottoman Fiscal Registers”. *Proceeding of V. International Conference on Social and Economic History of Turkey*, Ankara, 1990.

McGowan, B. “Food Supply and Taxation on the Middle Danube, 1568–1579”. *Archivum Ottomanicum*, 1969.

Marker, H.J. and K. Pagh (Eds.). *Yesterday*. Proceedings from the 6th AHC International Conference, Odense, 1994.

Mawdsley, E., N. Morgan, L. Richmond and R. Trainor (Eds.). *History and Computing III*. Manchester University Press, 1990.

Schürer, K. and S.J. Anderson. *A Guide to Historical Data Files Held in Machine-readable Form*. Association for History and Computing, 1992.

Simoudis, E., “Reality Check in Data Mining”. *IEEE Expert*, 1996, 26–33.

Singer, A. “The Countryside of Ramle in the Sixteenth Century: A Study of Villages with Computer Assistance”. *Journal of the Economic and Social History of the Orient*, 339(1) (1990), 59–79.

Wagstaff, J.M. *The Evolution of the Middle Eastern Landscapes*, 1985, pp. 190–204.