Analyzing a complaint database by means of a genetic-based data mining algorithm

J.J. van Dijk
R.S. Choenni
F.L. Leeuw

Content

Abstract

Although government organizations have recognized the potentials of data mining as a tool to extract interesting knowledge form their data sources, the application of data mining is still in its childhood. Partly, this is due to the fact that a data mining application requires sufficient knowledge of the database at hand, the domain to which the data refer, and the mining algorithms. In this paper, we have analyzed the complaint database of the National Ombudsman who handles complaints with regard to almost all government organizations in the Netherlands. The analysis aims to investigate the trends with regard to (the handling of) complaints during a period of 25 years. For this purpose, we have exploited a mix of conventional database techniques and a genetic-based mining algorithm. Our experience is that such a mix and the guidance by domain experts are necessary to produce useful and interesting results.

1 Introduction

Today, government organizations are putting a significant amount of efforts in the development and implementation of data sources, such as data warehouses. Although a wide range of data mining tools and algorithms can be found in the literature and the market-place, the successful application of data mining techniques on data sources in government organizations is still in its childhood (Bach, 2003). Partly, this is due to the fact that a significant amount of effort is required to extract useful knowledge from the data sources available in these organizations, since a good understanding of the data, the domain to which the data refer, and the mining algorithms, is required (Fayyad et al., 1996; Perng & Chang, 2004). Nevertheless, government organizations recognize the potentials of data mining as a tool to extract interesting knowledge. So far, the efforts (that are going on) to apply data mining in government organizations can be mainly found in the field of fraud detection and (counter)terrorism (Bonchi et al, 1999; Seifert, 2007; Brasch, 2005). In this paper, we focus on the application of data mining on complaints received by the National Ombudsman, which is a High Council of State in the Netherlands.

The National Ombudsman handles complaints regarding almost all government and a few private organizations related to the government (usually considered to be quangos). Each year, the Ombudsman receives thousands of complaints by letter and by e-mail. Telephone enquiries can usually be settled immediately, however in a number of cases these can also lead to written complaints. The handling of a written complaint begins with the creation of a (complaint) dossier. In 1985, the Ombudsman began to automate the processes in relation to these dossiers.

The 25-year anniversary of the National Ombudsman has marked a need to investigate the trends with regard to complaints and the handling of complaints during this period. This could be done in a conventional way by selecting and analyzing a random sample of dossiers from the dossiers that are available at the National Ombudsman or by devising questionnaires and distributing them to those who are familiar with the types of complaints that are submitted to the National Ombudsman and then test a few hypotheses on relationships between several variables (Jacobs, 1994; Timmer & Niemeijer, 1994). However, we have applied another approach that differs on two points from the conventional approach. We have used the data available in the electronic database of the National Ombudsman. One advantage is that analyzing the data is emphasized and not so much the carefully determining a random sample and the gathering of data. A disadvantage on the other hand is that, if specific data has been omitted, it is not easy to gather this and subsequently include it in the analysis. Secondly, we exploited a mix of conventional database and data mining techniques to gain a greater understanding of the complaints and to find trends with regard to complaints and the handling of complaints.

Our approach was feasible, since the National Ombudsman built up a relational database, referred to as *complaint database*, which contains information from over 140 thousand dossiers. The need to investigate trends leads in our approach to two sets of questions (see Section 2.2). The first set of questions can be answered by posing a set of SQL statements to the complaint database. In order to answer the

second set of questions we implemented a genetic algorithm and regarded the complaint database as a search space. Genetic algorithms are heuristic search strategies inspired by natural genetics and evolutionary principles (Goldberg & Holland, 1998). They have been proven to be promising in a wide variety of applications. We used the genetic algorithm to search for interesting classes within the database. A class can be regarded as conjunction of predicates. For example, the class of people complaining about the way they are treated by municipalities can be expressed as: *nature* ∈ ['treatment'] ∧ *type_of_organization* ∈ ['municipals'], in which the attributes *nature* and *type_of_organization* respectively describe the type of a complaint and the type of organization against which a complaint is filed.

Our approach yields a number of interesting results for the National Ombudsman. We have been able to determine interesting profiles within certain groups of complaints. Throughout the years, the number of complaints received by each implementing body has varied considerably. Our study has also revealed striking differences between the various types of intermediary organizations or agencies in terms of the successful handling of complaints. Sometimes, within a class it was worthwhile to apply conventional techniques to test predefined hypothesis defined by domain experts. For example, within the class of successful complainants, we tested whether repeat players are more successful in complaining than one-shotters. It appeared that repeat players are *not* more successful than one-shotters.

The remainder of the paper is organized as follows. In Section 2, we define our mining questions in detail and outline the database characteristics. In Section 3, we briefly discuss the genetic algorithm that we have used. Section 4 is devoted to the results that we have obtained. Finally, Section 5 concludes the paper.

2 The database and mining questions

Before we introduce the database and mining questions, we describe the complaint handling procedure in Section 2.1. After enumerating the (mining) questions that we have answered in Section 2.2, we discuss, in Section 2.3, the complaint database in more detail.

2.1   Complaint handling

If a member of the public is dissatisfied with the result of an internal complaints procedure by a government organization, or the way a complaint is handled, he has a period of one year to ask the National Ombudsman to investigate the case. The Ombudsman can deal with the request in two ways. The first is to launch a detailed case-investigation. First of all, his staff will try to assemble the facts regarding the 'incident'. They may carry out an on site inspection or interview the people involved by telephone or face to face. Questions will be put to the complainant, the government organization and the official concerned. They have an obligation to provide information. Both sides have an opportunity to comment on each other's answers. Through the operation of this 'right of reply' the National Ombudsman eventually arrives at a decision on the actions under investigation. The investigation may end with the publication of a report (with names removed). Reports of the National Ombudsman are in the public domain and may be brought to the attention of the media.

The second way of dealing with a complaint is to intervene directly to try to solve the complainant's problem. In this case, the National Ombudsman will contact the government organization and see whether there is any prospect of achieving a swift solution. If the result is satisfactory, the case will be closed by sending the complainant a letter explaining the outcome. This approach is used mainly when problems can easily be sorted out (e.g. by getting the organization to answer a letter) or when it is important to the complainant's well-being that the National Ombudsman should find a rapid solution to the problem.

In both cases, structured data with regard to the complaint are stored in the complaint database.

2.2   The mining questions

As noted in the foregoing, the Netherlands National Ombudsman is interested in how the nature and handling of complaints have evolved in a period of 25 years. To answer this question, we have divided it into a number of sub-questions. This has been done in cooperation with employees of the National Ombudsman, who are experts on handling complaints submitted to the National Ombudsman. Some of these questions can be answered by querying the complaint database by means of SQL statements, while others may be classified as mining questions. Compared to questions that can be formulated by means of SQL, mining questions have a high degree of

vagueness and incompleteness (Choenni et al., 2005). The following questions could be solved by querying the database by means of SQL statements:

— How has the number of complaints changed over the years? Is it possible to identify shifts in these trends?
— How are the complainants spread in geographical terms throughout the Netherlands? Is it possible to identify shifts in these trends?
— Which government organizations have received the most complaints? And what types of organizations are these? Is it possible to identify shifts in these trends?
— In order to make it possible to identify shifts in trends with regard to complaints in the questions mentioned above, the database has been divided into periods of 1 year and periods of 5 years. We have marked the following questions as data mining questions.
— What kinds of complaints can be marked as successful, i.e., have more than average chance to be declared (partly) founded, by the Ombudsman?
— What are profiles of interesting complaints?

To answer these questions, we have implemented a genetic-based algorithm and have applied it on the complaint database. We note that the answers on the data mining questions may be formulated as a conjunction of predicates and represents a subset of the database. Once an interesting subset is found, one may pose (SQL) queries on this subset. In the remainder of this paper, we will focus on the data mining questions and the application of the genetic algorithm. For those who are interested in the answers of all questions, we refer to Van Dijk et al. (2007).

2.3   The database

As noted before, the Ombudsman has a complaints database featuring information regarding complainants and complaints starting from the mid 80's. This complaints database is used to store structured information regarding dossiers and related matters such as the nature of the complaint, the date of receipt and date of handling of a complaint, the decision taken by the Ombudsman concerning the complaint, the intermediaries involved, etc. It is also recorded whether the complaint was submitted by a legal entity or a natural person. In many cases, address details are available in many cases. Personal information, such as age, gender or level of education, is not recorded in the database. The database is a relational database that consists of 21 attributes. The cardinality of the database is 317,244 tuples. The data is extracted from 143,847 dossiers, submitted by 115,099 applicants (also referred to as 'complainants'). The attributes of the database are described in Table A.1.

After determining the attributes that are available and useful to answer our mining question, we cleaned up the database, i.e., we checked for incorrect and inconsistent information. In general, the quality of the data in the database is quite good. There are not many missing values and the data in the database is consistent. Only the attribute 'area code' appeared to be unreliable, since almost 95% of the values of this attribute are missing values. However, we encountered the problem of data evolution for some attributes, i.e., the definition and meaning of data changed over time. The attributes 'type of organization', 'code of organization', and 'division' suffer from this problem. Government institutions have been merged and ministerial departments have also been converted into autonomous organizations. For example, two former

organizations, the Joint Administration Office and CADANS, merged into the Employee Insurance Schemes Implementing Body today. However, we still found the name of the former organizations in the database, although these organizations do not longer exist. Another example is that the former student loan company of the Ministry of Education & Science is nowadays an autonomous organization (a quango). We note that the problem of data evolution is not a typical problem of government organizations. We see the same phenomenon in different situations. For example, Kosovo declared it's independence from Serbia in early 2008, and therefore Serbia means – according to the Kosovo government – nowadays something else than in 2007.

As a consequence of data-evolution, the classifications of the complaints according to the government institution that is the subject of the complaint may result in inaccurate comparisons over a period of several years. We managed to resolve this problem by dividing the complaints into government institutions in their current form. We have added three additional attributes in the complaint database, describing whether an organization today can be classified as an implementing body or not, what the current description of an organization would be that does not longer exists, and to what kind of division a complaint would be pertained today if the recorded division in the database no longer exists. These attributes are listed in Table A.2. Due to the detailed method of storage in the complaints database, in the case of the present government institutions we were also able to classify the complaints of their (legal) predecessors where applicable.

Some initial results of the mining algorithm provoked us to the addition of some extra attributes in the database. For reasons of completeness of the mining database, we now discuss these attributes in this section. On the basis of our initial results, the National Ombudsman has recommended us to direct our search in the database also to repeat players and one-shotters (Galanter, 1974/'75; Jacobs, 1994). This characterization has been introduced in 1974 by Galanter[1]. Furthermore, from our initial results we have noted that the time required by the Ombudsman to deal with a complaint might become crucial in some profiles. To speed up our search, we added some extra attributes with regard to one-shotters/repeat players and the processing time required by Ombudsman to deal with a complaint. The description of these attributes is also listed in Table A.2.

---

[1]   M. Galanter: Why the 'Haves' Come Out Ahead: Speculations on the Limits of Legal Change. Law & Society Review, 1974.

3 Genetic-based algorithm

Many data mining problems can be formulated as search problems. A database is regarded as a set of tuples, and each (projected) subset of tuples is considered as an element in the search space. The problem is to select interesting subsets without inspecting the whole search space (Choenni, 2000). For example, the National Ombudsman may be interested in the profiles of successful complaints, i.e., complaints with (more than average) chances to be declared justly, can be modeled as a search problem. Consider the relation *Complaint*(*zip_code, place, nature, date_received, type_of_organization, type_of_intermediary, ind_intervention*), in which the attributes *zip_code* and *place* refer to the complainant's address, while the remaining attributes refer to the complaint. *Nature* is a classification of the subject matter of a complaint, *type_of_organization* is the organization to which a complaint pertains, and *type_of_intermediary* is the type of the intermediary who submits the complaints on behalf of the complainer. The attribute *ind_intervention* records, whether a complaint has been resolved by means of an intervention. This attribute is an indicator of whether a complaint can be considered as successful. Let us assume that a complaint that yields an intervention by the National Ombudsman is considered as successful. The set of tuples that satisfies to the predicate *ind_intervention* = 'yes' is called the target database.

The challenge is to select a conjunction of predicates that represents the class of successful complaints within the target database. Assume that this class is formed by complainants living in The Hague or Amsterdam, and who complain about the way they are treated by the municipalities. In this case, an expression like: *type_of_organization* ∈ ['municipals'] ∧ *nature* ∈ ['treatment'] ∧ *place* ∈ ['The Hague', 'Amsterdam'] should be searched for, which actually represents a projected subset of tuples in the database.

In general, the search spaces that should be inspected in order to answer mining questions are very large, making an exhaustive search infeasible. Therefore, heuristic search strategies are of vital importance to data mining. In the literature a wide variety of heuristic search strategies have been reported to walk efficiently through a search space, e.g., hill climbing, simulated annealing, genetic algorithms, etc. The success of an algorithm is often dependent on the structure of the search space. For example, a hill climber will generally perform better on a search that consists of a few optima than a genetic algorithm, while the latter will perform better if the search space consists of many optima. The reason is that a hill climber terminates if it reaches an optimum, while a genetic algorithm does not. In a search space that contains many optima, it will generally be worthwhile to continue searching the space at the first obtained optimum. Unfortunately, the search space that stems from the complaint database neither has a specific structure nor is the structure known in advance. Therefore, to tackle the mining questions that are listed in Section 2.2, we have adapted the genetic-based data mining algorithm reported in Choenni (2000). This algorithm does not pose any restrictions on the structure of a search space. Second, the problem of partitioning (numerical) attribute values, which is a tough problem in general, is solved by choosing a suitable manipulation operator.

The algorithm is characterized by the representation of individuals, a fitness function that evaluates an individual, and the manipulation operators cross-over and mutation. We represent an individual as an expression. This representation seamlessly fits in the field of databases. The fitness function takes care that extracted knowledge from the database is supported by a significant part of the database and that trivial knowledge is discarded beforehand. The mutation operator is implemented such that an expression undergoes a minor modification. The cross-over operator takes two expressions, selects a random point, and exchanges the sub expressions behind this point. A mutation operator is used for a local search, i.e., a search among neighbors for better expressions, while a cross-over operator causes a search in a complete different part of the database.

An expression that potentially represents an interesting group of tuples is called a class. We posed the following predefined criteria on a class. A class should cover about 70 to 80% of the tuples of the target database. We note that the target database is the part of the complaint database that is subjected to investigation. Furthermore, a target database with a few tuples is not allowed, since statements based on classes derived from these target databases are not reliable. Furthermore, a class should be described by at least two and at most five predicates. The rational behind this requirement is that we want to describe profiles by a small number of characteristics that cover a substantial part of the target database.

In the next section, we discuss the results that we obtained by applying the genetic-based algorithm.

4   Results

We have used the search algorithm to search for profiles of complaints in the database. We distinguished different target classes and varied the parameters of the genetic algorithm. For a discussion of the different parameters that can be varied in the algorithm, we refer to Choenni (2000). Our data mining exercise resulted in a number of potential complaint profiles. We submitted these profiles to domain experts at the Ombudsman, who then gave an indication of the parts of the database that would merit further exploration. Most of the results that have been found by our algorithm could be well interpreted by the domain experts. In the following we highlight some of the results that we have found.

To answer the mining question '*What kinds of complaints are marked as successful, i.e., more than average chance that the complainant proved to be (partly) right, by The Ombudsman?*' we defined as target class with the expression *decision* = NOT 'unfounded'. This expression selects the decisions in the database that are founded, partly founded or where no decision has been taken. Within the target class, we found the following expression: *nature* = 'long processing time' $\land$ *indication_of_intervention* = 'yes', which means that if the National Ombudsman will intervene in a complaint about a long processing time needed by a government organization, the complainant has good chances that he/she will be proved (partly) right. During the evaluation of the answers on this question provided by the algorithm, we noted that the type of intermediary also plays an important role. We note that the emergence of intermediaries is remarkable. With the arrival of the National Ombudsman, it was expected that the gap between members of the public and the government would be decreasing. However, an intermediary increases the gap between two parties (Niemeijer, 2007). Nevertheless, we were directed to question: what are the success rates of different type of intermediaries? We have defined success in different ways. We considered the decisions that reached in reports, and defined success in two ways, namely as *fully or partially founded* and *fully founded*. We obtained the following interesting table. We note that *reports* refers to the number of reports that has been yielded due to the complaints submitted by an intermediary.

Table 1   Success rates of different types of intermediaries.

Type of intermediary	Reports	Success rate (fully or partially founded)	Success rate (fully founded)
Association/foundation	133	86%	54%
Private (individual)	127	82%	51%
Legal aid agency / Legal Counter	299	79%	50%
Trade union	81	70%	48%
Citizen's advisers	62	84%	47%
Lawyer	1,590	82%	43%
Administrative office/tax consultant	352	65%	38%
No intermediary	9,333	73%	38%
General social services	90	74%	37%
Action group/interest group/information lines	73	84%	27%
Legal complaints office	213	74%	18%

If we consider the success rates of fully founded cases over the years, the 'best' intermediary has been an association or foundation (for example the Netherlands Refugee Foundation); 54% of the complaints submitted by this type of intermediary are declared founded by the Ombudsman. The second best type of intermediary are private individuals (51%) and the third most successful type of intermediary are the legal aid agencies/Legal Counters (50%). The relatively low success rate of lawyers in Table 1 is remarkable, since lawyers may be considered as professionals with a special interest and expertise in this field. The success rate of the lawyers may be partly declared by the fact that they also submit complaints to the National Ombudsman to obtain merely a decision, no matter what the outcome of the decision will be. These decisions might be used by lawyers as means of trying to speed up cases and/or influence them in a particular direction on obtaining jurisprudence.

It is often the case that 'part' of a complaint is declared founded whilst another part is declared unfounded. If we consider the data again from this point of view, associations and foundations still appear to be very successful (86%), however they are now followed by citizen's advisers and action groups etc. (84%). The third place is now shared by private individuals and lawyers (82%). Compared to the success rates for the fully founded category, the category of fully or partially founded cases is less distinguishing. However, the scale of the differences between the intermediaries is noticeable. Given the costs associated with collaborating with intermediaries and the increased use of intermediaries (Niemeijer, 2007; Van Velthoven & ter Voert, 2004), further research in this area is desirable.

To answer the mining question with regard to interesting profiles, we have distinguished different target classes such as complaints in which the Ombudsman has intervened, complaints on which the Ombudsman has produced a report, complaints submitted by complainants from the various provinces and complaints submitted by frequent complainants. However, we have not found any interesting expressions in these target classes.

We expected to find in some of the profiles that frequent complainants would be more successful than infrequent complainants, since frequent complainants have more experience with the procedures of the National Ombudsman. Therefore, we analyze the data to find out why our expectation did not come out. The results are presented in Table 2. It is obvious from the results that highly active complainants are not more successful than frequent or infrequent complainants. A possible declaration for this observation might be a topic for research.

Table 2   Success rate of different types of complainants.

Type of complainant	Successful (fully or partially)	Successful (fully)
Infrequent complainant	92%	79%
Frequent complainant	90%	75%
Highly active complainant	88%	68%

The results that we obtained with the mining algorithm as well as the results that we obtained on the first set of questions as presented in Section 2, provide the National Ombudsman a comprehensive insight into their complaint database. The answers on the first set of questions of Section 2 reveal amongst others that throughout the years,

the number of complaints received by each implementing body has varied considerably. Since a detailed discussion on the answer of this set of questions is beyond the scope of this paper, we refer to Van Dijk et al. (2007).

5   Conclusion

The Netherlands National Ombudsman is interested in how the nature and handling of complaints have evolved in a period of 25 years. To answer this question, we have analyzed their complaint database with conventional database techniques, i.e. posing SQL statements on the database, and a genetic-based data mining algorithm. We have split the above-mentioned broad question into a set of questions that could be handled with a conventional database technique and a set of questions that could be marked as mining questions. For the latter set of questions, we used a genetic-based data mining algorithm. We have observed that conventional database and data mining techniques are complementary tools for searching useful and interesting knowledge.

Furthermore, as expected, the role of domain experts appeared to be a key asset in directing the search in the database. In our case, the proper application of both types of techniques and the guidance of the domain experts, have led to a comprehensive insight into the complaint database and exposed some interesting results.

References

Bach, M.P. (2003). Data mining applications in public organizations. *Proceedings of the 25th International Conference on Information Technology Interfaces (ITI), 211-216.*

Bonchi, F., Gianotti F., Mainetto D., & Pedreschi, D. (1999). Using Data Mining Techniques in Fiscal Fraud Detection. *Proceedings of the 1st Conference on Data Warehousing and Knowledge Discovery (DaWaK)*, 369-376.

Brasch, W.M. (2005). Fool's Gold in the Nation's Data Mining Programs. *Social Science Computer Review, 2005.*

Choenni, R. (2000). Design and Implementation of a Genetic-Based Algorithm for Data Mining. *Proceedings of the 26th International Conference on Very Large Data Bases (VLDB)*, 33-42.

Choenni, R., Bakker, R., Blok, H.E., Laat, R. de (2005). Supporting Technologies for Knowledge Management. In W. Baets, *Knowledge Management and Management Learning: extending the horizons of knowledge-based management*, 89-112. United States: Springer.

Coleman, J.S. (1982). *A Symmetric Society*. United States: Syracuse University Press.

Dijk, J.J. van, Leeuw, F.L., & Choenni, R.S. (2007). Complaint profiles, success rates and intermediaries. In *Werken aan behoorlijkheid, Nationale ombudsman 2008* (pp. 297-326). The Hague, The Netherlands: Boom Juridische uitgevers.*

Fayyad, U.M. et al. (1996). *Advances in knowledge discovery and data mining*. Menlo Park, CA, United States: American Association for Artificial Intelligence.

Galanter, M. (1974/1975). Afterword: Explaining Litigation. *Law & Society Review, 9,* 345-368.

Goldberg, D.E., & Holland, J.H. (1998). *Genetic Algorithms and Machine Learning.* United States: Springer.

Jacobs, W. (1994). *Complainants at the National Ombudsman: one-shotters and repeat-players.* Gouda, The Netherlands: Quint bv.*

Niemeijer, E. (2007). *A world of disputes; about the use of legal and non-legal procedures.* The Hague, The Netherlands: Boom Juridische uitgevers.*

Perng, Y.-H., & Chang, C.-L. (2004). Data mining for government construction procurement. *Building Research and Information, 32*(4), 329-338.

Seifert, J.W. (2007). *Data Mining and Homeland Security: An Overview.* United States: CRS Report for Congress.

Timmer, J., & Niemeijer, B. (1994). *Citizen, government and National Ombudsman.* The Hague, The Netherlands: SDU Uitgevers.*

Velthoven, B. van, & Voert, M. ter (2004). *Paths to Justice in dispute resolutions (in Dutch: Geschilbeslechtingsdelta) 2003.* The Netherlands: WODC.*

We note that references ending with a star (*) are in Dutch.

Appendix 1   Attribute lists

Table A.1   Attributes of the complaint database

Characteristics	Description
Key word	Key word to describe the complaint. This may be a general key word or a specific key word relating to the organization that is the subject of the complaint.
Extralegal indication	Indication of whether the complaint is extralegal.
Termination	Reason for termination (e.g. investigation concluded, second chance imposed).
Decision	Decision in the report: founded, unfounded, no decision or a combination of these.
Date of receipt	Date on which the application was received by the Ombudsman.
Date of completion	Date of the last document which brings the application to a conclusion.
Method of submission	Shows whether the complaint was received electronically or in writing.
Nature of the complaint	A classification of the subject matter of the complaint. E.g. long processing time
Indication of recommendation	Indication of whether a recommendation has been attached to the report.
Indication of suitability for publicity	Indication of whether the investigation is suitable for publicity purposes.
Authority	Power of investigation (usually whether the Ombudsman is or is not obliged to commence an investigation)
Type of organization, code of organization, division	Three characteristics that jointly describe the government organization that is the subject of the complaint.
Assessment	The Ombudsman's assessment of the organization.
Area code (phone number)	The area code of the complainant's landline number.
Postcode	Postcode of the complainant's residential address.
Place	The complainant's place of residence.
Province	The province in which the complainant resides.
Type of complainant	Legal entity or natural person.
Type of intermediary	For example: lawyer, relatives of complainant, etc.

Table A.2   Attributes added to the complaint database

Derived characteristic	Description
Indication of implementing body	Indication of whether the organization is classified as an implementing body today.
Oganization_Name	Description of the organization in its current form.
Nature_Division	Kind of division to which the complaint would pertain today.
Duration of processing	The date of completion minus the date of receipt with regard to a submitted complaint.
Indication of frequent complainant	Indication of whether a complaint is at least the fifth complaint submitted by the complainant.
Indication of highly active complainant	Indication of whether a complaint is at least the tenth complaint submitted by a complainant.
Indication of intervention	Indication of whether a complaint has been resolved by means of an intervention.