DataDetective product outline

April, 2008

Introduction

DataDetective, Sentient's own software suite for *matching* and *mining*, enables organizations to benefit from the enormous value of their data. Users are able to gain insight easily, build models, detect patterns and predict trends, by applying cutting-edge technology from the field of Artificial Intelligence built in DataDetective. Key qualities are: powerful yet easy to use, scalability and flexibility.

Since **1991**, DataDetective has continuously been improved and applied in a wide variety of areas - To name a few:

- The **Dutch police** uses DataDetective since 2001 to match and mine crime databases in a national project to implement data mining in the police organization.

- Insurance firm **Delta Lloyd** uses DataDetective to analyze customer databases to understand market trends, recognize opportunities and improve enterprise-wide customer intelligence.

- **CDR**, the largest music library in Europe uses DataDetective to recommend artists and albums to their customers. See www.muziekweb.nl (section `muziekadvies')

- The **Dutch public library association** applies DataDetective as the main book and author recommendation service on the web and in all public libraries. See romanadvies2.bibliotheek.nl

- DataDetective enables **marketing professionals** in several businesses to gain insight in the various groups of customers and find the optimal channels to reach them.

- DataDetective is being applied to support **medical decisions**, for example to advice about the transfer op patients from one type of care to another. An important advantage for this application area is the ability of DataDetective to explain decisions.

- Using DataDetective, our consultants enrich customer databases with external survey data, using our **data fusion** technology.

- **Inspectors** use DataDetective to find patterns in websites that have been collected and categorized by the internet text tools of sister company Parabots.

- DataDetective has been used as a powerful **match engine** in many areas: job intermediation, witness descriptions and dating.

Flexible data interface

Use your favorite DBMS - With DataDetective, you have the option to build the data warehouse in your own corporate database. That way you can manipulate data in a familiar way and utilize investments made in database technology. DataDetective automatically transmits the changes to its own database (the *index*), which is required for optimal performance of the matching and mining algorithms. The only requirement for the database is that it supports ODBC (standard for almost every database nowadays).

Integrated ETL – DataDetective comes with an integrated ETL tool *DataStager* to manage the data warehouse. This tool takes care of all required steps in data preprocessing: Extraction, Transformation and Load, including scheduled downloads, data cleansing, automated data transformation, aggregation and automatic meta data specification.

Work directly with your rich data structures – DataDetective supports complex data structures, so the pre-processing normally required to make data suitable for data mining is

minimal. This reduces the costs of data mining and retains the quality and richness of the data. Many data types are allowed, such as: categories, scales, relational/hierarchical structures, and free text fields. Records can contain an unlimited number of variables and variables an unlimited number of values. Our most widely used DataDetective database has over 10,000 variables. The user interface is designed to work with such extensive data collections by using customizable screens, variable masks, value filters and various search functions. Code tables can be specified with D-designer to map a database representation to values and to user interface labels.

DataDetective also supports *object weights* with which certain records can be used to represent a specific number of other records (e.g. people), which is especially useful for survey data. These weights are handled transparently in DataDetective – it appears as if you are working with a database of the entire population while actually it is just a sample.

Fast

The DataDetective mining engine is extremely fast for several reasons:

- The DataDetective data format is proprietary and completely optimized for mining tasks while conventional operations are performed by the relational database.

- During the last ten years considerable effort has been put into optimizing the C++ core engine which features for example intelligent dynamic RAM caching and shared RAM usage by memory-mapping.

- Query materialization: if necessary, query results are stored physically to increase performance.

- DataDetective server is scalable by utilizing multiple machines and processors in realizing a data mining back office.

- DataDetective 2.1 will include near-trees to dramatically increase match speed.

Powerful user interface

Start data mining right away – Through the years, many different types of users helped shaping DataDetective's user interface into a user-friendly, intuitive design based on Microsoft Office look and feel, containing many intelligent agents for assistance and focused on tasks, not on techniques. All operations are initiated from the so-called 'organizer', the central repository of data and reports, which is easier and more advanced than a file system. Results are visually connected to the data. When performing operations, you specify the goal and DataDetective takes care of all the technical details under the hood. Algorithms are tuned automatically by agents in dialog with you. Decisions you make are monitored and warnings/suggestions are given when needed (e.g. -'The train set is too small…', -'Many missing values…').

Have a dialog with your data - DataDetective functions are tightly integrated, which allows you to explore data step by step. For example: 1) segment a dataset with D-segment, 2) analyze a segment with D-profile, 3) save the cluster as a dataset in the organizer, 4) use the dataset to profile against another dataset, 5) take a random sample of the dataset and use it to train a prediction model, 6) use the rest of the dataset to test the prediction model. Or: 1) analyze the buyers of product X using D-profile, 2) select characteristics from that profile and use them in D-query to find similar records, 3) exclude current customers from the resulting dataset, 4) D-profile the result to study potential customers for product X.

DataDetective offers the basic data analysis functions, such as data browsing, searching, a query editor, frequency charts, scatter plots, descriptives, and correlations. Besides these features, DataDetective has many unique functions. The most prominent modules are discussed below: D-segment, D-query, D-profile, D-model, D-map, D-designer and D-server.

D-Segment

Portray your data – Probably the most eye-catching feature of DataDetective is *Looking Glass*- the animated clustering tool. In its dynamic process, records move around in two dimensions until the picture becomes an optimal projection of all data dimensions. The result is an interactive visualization of coherent groups and the relation between these groups. Apart from identifying segments, Looking Glass visualizes the shape and the similarity of these segments. A segment can be a coherent blob, but also a band ranging from one prototype of customer to another, for example from early adaptor to late adaptor.
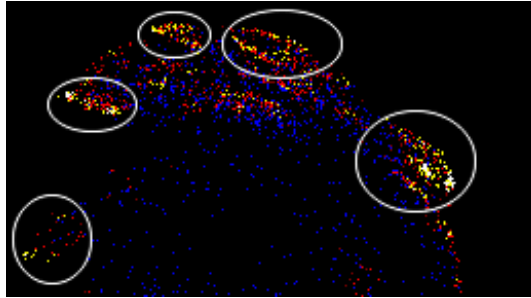
You can explore the resulting segmentation by zooming in, colorizing variables or intersections, comparing clusters using D-profile and naming them. The unique algorithm and visualizations are original developments of Sentient.

The Looking Glass algorithm is NOT based on principal component analysis. Instead of finding linear combinations of dimensions, the entire multidimensional space is deformed and projected to make the representation as perceptively accurate as possible within 2 dimensions. Furthermore, Looking Glass is less susceptible to noise and missing values.

Application examples for Looking Glass:

− Finding customer groups that need different tone-of-voice and communication channels. Get to know different types of customers, instead of a single customer profile.

− Discovering anomalies by investigating small groups outside the mainstream segments. For example: money transactions, internet behavior.

− Detecting patterns of behavior to spot for example series of crimes with a similar M.O.

D-Query

Select records by similarity – Using the D-Query module you can specify fuzzy queries (based on similarity), as well as conventional queries (based on yes/no criteria). Fuzzy search retrieves records that –resemble- a description. The criteria are matched with all records in the database and the best matching records are returned, ordered by similarity. The fuzzy search algorithm supports bi-directional matching (which is necessary for job matching for example) and batch processing of match jobs.

Advantages of fuzzy search:

− Easy: no interactive construction of queries to tailor the results. Simply specify what the ideal outcome should look like.

− Efficient: the results are ordered by similarity: best first, which saves time in surveying the results.

− Precise: all data can be used; even uncertain data. Criteria can be made more important by increasing the weight.

− Robust: insensitive to noise (e.g. errors). Individual values may differ, as long as the other values are similar. Example: a man disguised as a woman is still retrieved from a police database based on a witness report because many features are similar, even though gender is completely wrong.

− Fast: The DataDetective engine is optimized for fuzzy search. A similar operation using a state-of-the-art DBMS takes 400 times longer.

Applications for fuzzy search:

− Matching supply and demand (E-markets), such as second hand cars, real estate, blind dates, and jobs.

− Find the needle in a haystack; search a criminal database with an eye witness report.

− Specify a group using a prototype: a market analyst can define a target group of prospects that resemble the ideal customer. A database normally contains just a few *ideal* customers, whereas a much larger and representative number closely resemble the description.

The diagram below shows how D-query selects objects by similarity (height 185, age 27) while applying conventional criteria at the same time (height>172). It also illustrates that a conventional query around 185 and 27 has no result at all.
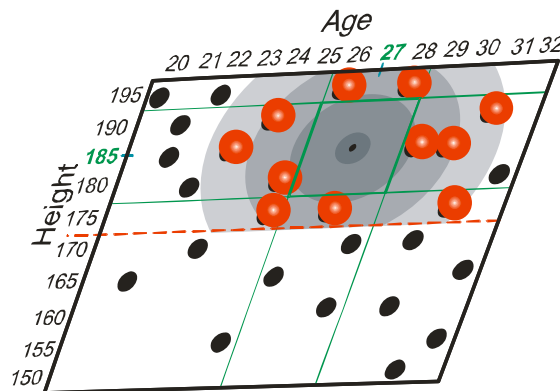
D-Map

Map your data on the real world - The D-Map module offers many GIS (Geographical Information System) functions to visualize data on geographical maps. Aggregate data (e.g. sum, average, density) is displayed per region in color, pie chart or bar chart. The D-Map visualization features scrolling, zooming, controlling layers, and changing the region detail. Again, this function is tightly integrated with the rest of DataDetective - by clicking on a region you can quickly make a D-profile, or chart a variable.

DataDatective standard contains a basic geographic module. For more advanced visualization there is a tight integration with Mapinfo[tm] – with just one click of the button you can create insightful geographical visualizations, such as hotspots or contour graphs.
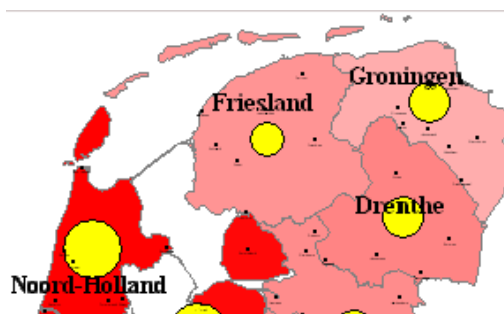
D-Profile

Get to know your data with one click – Finding patterns in data is normally based on creating a theory and testing it on the data (e.g. 'Customers of product P are relatively older than our average customer'). DataDetective enables you to find –all- the relevant characteristics of a group with just one button click. The result is a list of characteristics, ordered by distinction. Use it for example to profile top customers, product portfolios or risk groups. Nothing is overlooked.

This profiling function works for all supported data types, including categories, scales, multivalues and text fields. A D-profile results in a highly interactive report allowing you to drill-down, zoom-in, make charts, sort and filter characteristics, etcetera.

Work with real-life datasets without oversimplification - In data analysis, people are typically divided into groups where they are either members of the group or they are not. There is no in-between, which often is a simplification of modern reality in which people are not easy to categorize at all. DataDetective addresses this issue by providing the *Fuzzy set*, of which records are members to certain degrees. When such a fuzzy set is analyzed, all statistics are automatically corrected by the degree of membership. For example: a fuzzy set 'New York times reader' is specified by setting the membership to the number of New York times a person reads a month. Someone who reads them all is very much a New York Times reader, while someone that reads just 1 is a moderate reader. When analyzing such a set, to create for example an age histogram, the 100% readers have much more weight than the moderate readers. In other words: instead of a yes/no relation to a concept, people have a continuous relation, which is more precise and allows you to analyze more data and get better results.

Decision trees – in order to understand a specific target group (e.g. 'Top customer), it can be analyzed in a decision tree. Such a tree uses D-profile to find the most important characteristic of the group (e.g. 'Lives in the south') and splits up the set accordingly. Next, each subgroup can again be analyzed and split. This way the tree could for example indicate that a large part of the top customers is male, lives in the south and has account manager X.

D-Model

Build prediction models in just a few steps - With D-model you can build models to estimate/predict for example direct mail response, customer behavior, potential turnover, medical risks, or the weather forecast. In five easy steps you are guided in selecting an algorithm, defining train set and test set, analyze the performance using gains charts, etcetera. Each algorithm included in DataDetective has its pros and cons. Some algorithms execute very fast but cannot offer an explanation like other algorithms can. One of the unique algorithms in DataDetective *RFC*, offers three interesting advantages: it can handle all the complex data structures supported by DataDetective (e.g. text, many variables), it has the ability to explain estimations, and it is able to estimate many variables at the same time. Another special feature of this algorithm is its ability to recommend products to customers using a unique algorithm to balance between products that are popular in general and products that are especially popular within customers segments.

D-Designer

Use or build vertical solutions –Vertical solutions can be created straightforwardly by configuring the user interface and data warehouse with the integrated development tool *D-Designer.* In case programming is required, specific extensions can be built using DataDetective's extensive plug-in architecture that allows bespoke software to tightly integrate with the application. Many of such applications have been built since 1992, which contributed to the flexibility and expandability of the DataDetective development environment.

D-designer allows the advanced user to import or link databases and define application-specific elements, such as data layout, field types, code tables, algorithm parameters and the level of advanced functionality. An expert system is available to assists in the process. With D-designer, DataDetective can be tailored to all kinds of users from different fields and with varying computer/data mining experience.

We offer a number of shrink-wrapped vertical solutions:

- DataDetective client monitor, for estimating customer potential and analyzing the effects on the total customer 'pyramid' for upcoming periods.

- DataDetective packaged with leading Dutch and European consumer surveys, such as NOM. See www.doelgroepdetector.nl.

- DataDetective-AHS, a law enforcement application for finding suspects by eyewitness reports, including multimedia support (e.g. portraits). The Dutch police have acquired DataDetective-AHS for national use.

D-server

Deploy DataDetective in the back office - DataDetective is also available as a scalable server solution with an extensive public interface to perform matching & mining or manipulate the data/application design.

The server interface is dual: one interface is based on the portable SOAP standard (easy to create a web service); another interface is built using DCOM (easy integration with Windows development tools). Scalability is achieved by DataDetective's proprietary load balancing system, the Sentient Task Manager (STM), allowing developers to deploy DataDetective server over multiple processors and machines. In addition it takes care of fault tolerance and problem management. DataDetective with STM is used by several system integrators in mission-critical applications, such as job intermediation and automated medical advice.

Sentient Information Systems
Singel 160
1015 AH
Amsterdam

info@sentient.nl
http://www.sentient.nl

tel: +31 20 5300 330
fax: +31 20 5300 331