On exploration and mining of data in education practice

Dr. Erik-Jan van der Linden (MagnaView)
Ir. Thomas Lammers (MagnaView; TU/E)
Ir. Martijn Wijffelaars (MagnaView)

EDM conference
Eindhoven, July 6, 2011
Speakers

Van der Linden, Lammers, Wijffelaars
MagnaView (company)

Our customers solve their business issues because we enable them to get the most out of their data, fast.

Speakers

Partly University Eindhoven spin-off

Lost for academia..?
Yes... Please interpret company blabla as enthousiasm

No... R&D ‘intensive’: 10.000 hours per year. Examples today.
Credentials: scientific publications
Cooperation Eindhoven TU - MagnaView

IEEE Visualization, InfoVis 2006
Baltimore, Maryland

UDC Seminar 2007
The Hague, the Netherlands

EuroVis 2008
Eindhoven, the Netherlands

Overview

A. Education practice: Context
B. Brief introduction MagnaView
C. Visual exploration in education
D. Data mining meets visualization
A. Education practice: Context

- Recognition of knowledge and innovative power as the drivers for economic development.
- Education is high on the political agenda.
- OECD Programme for International Student Assessment (PISA) considered important by politicians: how does our national sector do? → translate to policies which apply to institutions.
Context (ii)

- Insight of governments and politicians → Impossible to adapt system on basis of centralized regulations etc. → schools run education.
- Netherlands' conclusion: lump sum funding → right to spend = obligation to justify!
  Both spending and outcome.

Context (iii)

- N.B. International?
- Germany: visible shift. Ministeries and municipalities carry over budgets to schools
- Belgium: slowly shifting attention towards teachers
- Finland: best system in the world because of professional status of teachers; why change..?
Context (iv)

- Other development: data accumulates in school information systems.
- *If I were Coca Cola...*
- Data has become available → politicians: ‘if they HAVE the data, we can ASK!’
- And Calvinist Dutch: ‘we can publish...’
  - ‘Kwaliteitskaart’
  - ‘Trouw’

Context (v)

- So, external pressure on schools to USE their data for *external* reporting...
- ... but as a consequence use for *internal* improvement.
- Impossible to change the school on basis of centralized regulations etc. → individual teachers teach/individual student learn
- Of course, the one-on-one communication between teacher and student is key.
This context....

... is our playground...

Companies only survive if they start from ‘needs’ of their customers...

... which is different from...

EDM (2011 preface)

Data-driven/curiosity-driven.

The increase of e-learning resources such as interactive learning environments, learning management systems, intelligent tutoring systems, and hypermedia systems, as well as the establishment of state databases of student test scores, has created large repositories of data that can be explored to understand how students learn.

MagnaView: issue-driven
B. Brief introduction to MagnaView

About MagnaView

- Started in 2004 in Eindhoven - The Netherlands

- Our mission
  
  Our customers solve their business issues because we enable them to get the most out of their data, fast.

- What we offer
  
  α General purpose Visual Analytics Software
  
  β Turn key Analytics Products for specific sectors
  
  γ Full Service
**α General purpose visual analytics software**

<table>
<thead>
<tr>
<th>Tell me here...</th>
<th>Web-based software for end user works on laptop, PC, Smartboard, Tablets, iPhone, iPad, Blackberry</th>
</tr>
</thead>
<tbody>
<tr>
<td>.. and now (I really mean NOW) ...</td>
<td>Fastest tool in the marketplace</td>
</tr>
<tr>
<td>.. what really matters...</td>
<td>Visualize anything and easy to understand for everybody → create sense of urgency</td>
</tr>
<tr>
<td>... allow me to gain insight....</td>
<td>Easy mouse-click navigation to follow train of thought</td>
</tr>
<tr>
<td>... and to decide where to act.</td>
<td>Any level of detail available: in-memory analysis of large datasets</td>
</tr>
</tbody>
</table>

**α General purpose visual analytics software**

- **Desktop products**
- **Enterprise products**
- **OEM**
- **Cloud / SaaS Managed services**
**β Turnkey Products**

<table>
<thead>
<tr>
<th>We require information relevant for our sector...</th>
<th>MagnaView offers Products for specific sectors</th>
</tr>
</thead>
<tbody>
<tr>
<td>... and we really mean relevant.</td>
<td>Users are continuously involved in development of the MagnaView Products</td>
</tr>
<tr>
<td>We don’t want to bother about data</td>
<td>MagnaView offers truly turnkey products which include queries, data handling, etc</td>
</tr>
<tr>
<td>We want it integrated in our working environment</td>
<td>MagnaView cooperates with your current software vendor</td>
</tr>
</tbody>
</table>

**Health care:** MagnaView Microlab - Pathos
**β Turnkey Products**

*Legal:* MagnaView Patronus - Daan  
*Education:* MagnaView Cum Laude

---

**γ Full service**

<table>
<thead>
<tr>
<th>Want</th>
<th>MagnaView does</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fast installation and integration</td>
<td>takes care of all aspects of installation and integration</td>
</tr>
<tr>
<td>Direct results</td>
<td>manages full, agile implementation of projects</td>
</tr>
<tr>
<td>Lower costs</td>
<td>effectively train your expert users, BI/MI staff</td>
</tr>
<tr>
<td>Reports and analyses</td>
<td>The MagnaView consultants create great reports in no time</td>
</tr>
<tr>
<td>Maximize ROI</td>
<td>The MagnaView experienced consultants make your organization really profit from its data treasures</td>
</tr>
</tbody>
</table>
# Examples

## Profitability in legal firms - 1

<table>
<thead>
<tr>
<th>Company</th>
<th>Total</th>
<th>Total Profit</th>
<th>Map of Product</th>
<th>Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>Product 12</td>
<td>€ 1,174</td>
<td>€ 63,273</td>
<td>€ 19,063</td>
<td>4</td>
</tr>
<tr>
<td>Product 5</td>
<td>€ 2,387</td>
<td>€ 11,945</td>
<td>€ 5,790</td>
<td>3</td>
</tr>
<tr>
<td>Product 0</td>
<td>€ 21,540</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Product 6</td>
<td>€ 1,324</td>
<td>€ 17,515</td>
<td>€ 14,089</td>
<td>2</td>
</tr>
<tr>
<td>Product 7</td>
<td>€ 1,045</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Product 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Product 15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Product 11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Real Estate</td>
<td>Product 3</td>
<td>€ 7,145</td>
<td>€ 480,377</td>
<td>3</td>
</tr>
<tr>
<td>Product 4</td>
<td>€ 271,297</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Product 2</td>
<td>€ 180,397</td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

€ 430,000 € 490,310
Example: hospital (i). Starts from an advanced technique: *Process mining*. Wil van der Aalst at TU/E

‘Wiring scheme!’; ‘Spaghetti!’

Example: hospital (ii)
Change…

- the way the **doctor** thinks and acts!!

- the way the **teacher** thinks and acts...?
Cum Laude

Tool for management information

What?

- In use by 25 schools
- 75,000 pupils; 100,000 by the end of 2011
- 6-10 years of data
- Secondary Education
- Netherlands (partner in Germany)
- Web-based SAAS-solution
What?

- Information about
  - Population. Pupils and teachers
  - Absence
  - ‘Throughput’
  - Grades
  - ‘Kwaliteitskaart’

Cum Laude

Demo
Actual use of Cum Laude

- **Longer term**
  - Evaluation of throughput
  - Regular evaluations of teaching staff
  - Communication with primary education

- **Short term**
  - Absence
  - Grades
  - Meetings of sections
  - Ad hoc information/analyses

---

Actual use of Cum Laude

P.S. MagnaView is interested in contacts with Education Software Vendors for Cum Laude

martijn@magnaview.com

Tips are rewarded with free licenses ;-)
D. Data mining meets visualization

Follow the leader!

Data mining meets visualization

MSc Thesis Thomas Lammers. Now @MagnaView.

Not yet applied in schools, but show how general idea may apply to education data.


Original starting point...
Which routes do my customers take?
Starting observations

- Strength of visual exploration:
  - The user can easily input domain knowledge by creating the best possible visual representation, choosing the attributes he/she is interested in.
- Weakness of purely visual exploration:
  - User has to do all pattern recognition and grouping himself.
- Strength of data mining:
  - Uncover these patterns, groups, etc.
- Weakness of data mining...

Typical approach
But...

- Users with domain knowledge typically do not have understanding of the way DM-techniques/algorithms work
- DM is not going to be applied in practice (the MagnaView context)
- Is there another way to combine DM and Viz?

Previous work?

- There is little work on good combinations of DM and visualization.
- ‘Visual Analytics’ makes the claim, but the emphasis is on interaction and navigation. DM-techniques hardly found.
Quite some visual examples in EDM!

But...

- Not intended to serve as interface for user
- Exceptions: Brick and EDM Vis Tool...?
Observations (i)

- Grouping is an *incremental* activity. Users work in cycles
- Know it when they see it
- Browsing data can lead to discovery which groups are desired
- It is easier to critique than to construct.

Can we support that?

Observations (ii)

- *Any* group can be a real group for a user.
  - *Women, Fire and Dangerous things* (Lakoff).
  - How should we support that?

- When a user creates a visual representation, he/she in fact specifies what might be an interesting way to group data.
  - Could we use that?
Interactive grouping (i)

1. Have the user interactively change the current grouping:
   - Splitting all ungrouped entities into multiple groups using a clustering algorithm.
   - Splitting an existing group using a clustering algorithm.
   - Merging existing groups.
   - Creating a new group.
   - Moving an entity from one group to another.
   - Reducing the number of entities in a group.
   - Increasing the number of entities in a group.

Interactive grouping (ii)

2. Automatically re-group all entities, using NN classification algorithm.
3. If the current grouping is not the intended one, go to the first step.
DM techniques used

- Off the shelf; no intention to invent new DM-techniques.
- For clustering: K-means.
- Re-group: Nearest Neighbor

Interactive grouping
Leaders and Followers

- Metaphor
- Leader is example, that is put manually in a group
- Followers follow leaders

Clustering data demo (Iris dataset from the UCI Machine Learning Repository)

- Start Demo

<table>
<thead>
<tr>
<th>Width</th>
<th>Height</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.2</td>
<td>3.6</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>6.9</td>
<td>3.1</td>
<td>Iris-versicolor</td>
</tr>
<tr>
<td>5.0</td>
<td>3.2</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>5.9</td>
<td>3.0</td>
<td>Iris-virginica</td>
</tr>
<tr>
<td>5.1</td>
<td>3.5</td>
<td>Iris-setosa</td>
</tr>
</tbody>
</table>
Grouping: summary

- Group manipulation
  - New group
  - Delete group
  - Merge group
  - Rename group

- Leader manipulation
  - Promote to leader
  - Move leader
  - Demote leader

Split group
Change group Size

Similarity/Distance

- Data is labelled using Nearest neighbour; leaders are seed points.
Distance functions are *derived*

- Information in the view created by the user:
  Attributes used to group data.
  Attribute used for color.
  Attribute use for size.

The attributes for X and Y-position were used, other attributes, such as **Speed**, The **shopping cart number** or **time of the day** were not visualized, and thus not used in our **derived** distance function.
User test

Dislikes
- Resize group
- Change order

Likes
- Dragging data
- Lead-Follow Metaphor

Application to education data

A most disappointing demo
Conclusions

• Interactive grouping is promising
• We are just starting with applications in education domain.
• Applications should help solve business issues
  • Prevent pupils from dropping out...?
  • Prevent pupils from doubling classes...?
  • From choosing the wrong program...?
Carlijn

Sasha  Ilja