Practical Demos of Text Mining and Data Mining using SQL Server 2012

Mark Tabladillo Ph.D.   @marktabnet
Data Mining Scientist, MarkTab Consulting
Blog: http://marktab.net
Data Mining with SQL Server 2012

- SQL Server 2012 provides enterprise-level tools for advanced analytics.
- Microsoft was the first vendor to put analytics at the database.
- This release continues Microsoft's analytics leadership.
- MarkTab Presenting at PASS Summit 2012: [http://marktab.net/datamining/2012/06/19/presenting-pass-summit-november-2012/](http://marktab.net/datamining/2012/06/19/presenting-pass-summit-november-2012/)
Objectives

• Provide an introduction to data mining with SQL Server 2012
• Provide a data mining demo (previewing a 3-hour advanced workshop at PASS Summit)
• Provide an introduction to semantic search with SQL Server 2012
• Provide a text mining demo (previewing an intermediate session presentation at PASS Summit)
INTRODUCTION TO DATA MINING
Future: Data Mining is Deeply Cultural

- **Smartphones**
  - Windows Phone 8
  - iPhone (next)
- **Tablets**
  - Windows Surface
  - Kindle Fire HD
- **Search Engines**
  - Bing
  - DuckDuckGo
  - Google
  - Facebook
## Definitions

<table>
<thead>
<tr>
<th>Phrase</th>
<th>Goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Data Mining”</td>
<td>Inform actionable decisions</td>
</tr>
<tr>
<td>“Machine Learning”</td>
<td>Determine best performing algorithm</td>
</tr>
<tr>
<td>Query</td>
<td>Typical Result</td>
</tr>
<tr>
<td>-------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>T-SQL</td>
<td>Exact values and calculations</td>
</tr>
<tr>
<td>MDX</td>
<td>Exact values and calculations</td>
</tr>
<tr>
<td>DAX</td>
<td>Exact values and calculations</td>
</tr>
<tr>
<td>DMX</td>
<td>Values plus probabilities</td>
</tr>
</tbody>
</table>
What Enterprise Tools support Data Mining?

SQL Server Management Studio (SSMS)
SQL Server Data Tools (SSDT) [Formerly: Business Intelligence Development Studio (BIDS)]
  - SQL Server Integration Services (SSIS)
PowerShell version 2
What about the Competition?

• I have worked over the years with many tools
  • SAS, SPSS, R, Minitab, Stata

• I advise my clients to know the features and tradeoffs of all available software tools
  • Not Either/Or – but Both/And

• A 2012 Survey names Excel as the number-one commercial tool for analytics, data mining, and big data
  • http://marktab.net/datamining/2012/06/15/excel-number-commercial-tool-analytics-data-mining-big-data/
What Enterprise Tools support Microsoft Data Mining?

- SSMS
- SSIS
- PowerShell
How Microsoft Data Mining Works
### Structure

<table>
<thead>
<tr>
<th>Variable</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discretized</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
</tr>
<tr>
<td>Discretized</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
</tr>
<tr>
<td>Continuous</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
</tr>
<tr>
<td>Discrete</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
</tr>
<tr>
<td>Variable</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>------------------</td>
<td>------------</td>
<td>------------</td>
<td>------------</td>
<td>------------</td>
<td>------------</td>
<td>------------</td>
<td>------------</td>
<td>------------</td>
</tr>
<tr>
<td>Discretized</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
</tr>
<tr>
<td>Discretized</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
</tr>
<tr>
<td>Continuous</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
</tr>
<tr>
<td>Discrete</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
</tr>
<tr>
<td>Variable</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>---------------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Discretized</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
</tr>
<tr>
<td>Discretized</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
</tr>
<tr>
<td>Continuous</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
</tr>
<tr>
<td>Discrete</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
</tr>
</tbody>
</table>
### Structure

<table>
<thead>
<tr>
<th>Variable</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discretized</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
</tr>
<tr>
<td>Discretized</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
</tr>
<tr>
<td>Continuous</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
</tr>
<tr>
<td>Discrete</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
<td>Model</td>
</tr>
</tbody>
</table>

### New Sample

### Probabilities
Data Mining References

Data Mining

Website
- [http://marktab.net](http://marktab.net)
DEMO: DATA MINING
INTRODUCTION TO SEMANTIC SEARCH
Future: Most data is Text

Two Research Types

• Quantitative research = data mining
• Qualitative research = text mining

The future is combining both
Statistical Semantic Search

Comprises some aspects of text mining
Identifies statistically relevant key phrases
Based on these phrases, can identify (by score) similar documents
FileTables

Built on existing SQL Server FILESTREAM technology

Files and documents

- Stored in special tables in SQL Server
- Accessed if they were stored in the file system
Full-Text Search Enhancements

Property search: search on tagged properties (such as author or title)
Customizable NEAR: find words or phrases close to one another
New Word Breakers and Stemmers (for many languages)
From Documents to Output

Rowset Output with Scores

Office

VarChar

NVarchar

PDF
“Beyond Relational” vs. “Adoption”

Start with unstructured (meaning non-relational) data

Use Windows technology

- Reading and Writing Files (Win32 API)
- iFilters for reading proprietary formats

Develop indexed structure from unstructured data
(iFilter Required)

Documents → iFilters

- Full-Text Keyword Index “FTI”
- Semantic Key Phrase Index – Tag Index “TI”

Semantic Document Similarity Index “DSI” → Semantic Database
Languages Currently Supported

Traditional Chinese
German
English
French
Italian
Brazilian
Russian
Swedish
Simplified Chinese
British English
Portuguese
Chinese (Hong Kong SAR, PRC)
Spanish
Chinese (Singapore)
Chinese (Macau SAR)
Phases of Semantic Indexing

- Full Text Keyword Index “FTI”
- Semantic Key Phrase Index – Tag Index “TI”
- Semantic Document Similarity Index “DSI”

Phase 1

Phase 2

Integrated Full Text Search (iFTS)

Improved Performance and Scale:

- Scale-up to 350M documents for storage and search
- iFTS query performance 7-10 times faster than in SQL Server 2008
- Worst-case iFTS query response times less than 3 sec for corpus
- Similar or better than main database search competitors

(2012, Michael Rys, Microsoft)
Linear Scale of FTI/TI/DSI

First known linearly scaling end-to-end Search and Semantic product in the industry

Time in Seconds vs. Number of Documents
(2011 – K. Mukerjee, T. Porter, S. Gherman – Microsoft)
Text Mining References

Video

Semantic Search (Books Online) – explains the demo

Paper
DEMO: SEMANTIC SEARCH
Conclusion

• SQL Server 2012 provides enterprise-level data mining and text mining
• MarkTab Presenting at PASS Summit 2012: http://marktab.net/datamining/2012/06/19/presenting-pass-summit-november-2012/
• Resources:
  • Data Mining: http://marktab.net
  • Data Mining Blog: http://marktab.net/datamining
Questions?
Thank You for Attending