Knowledge Discovery for Business Intelligence

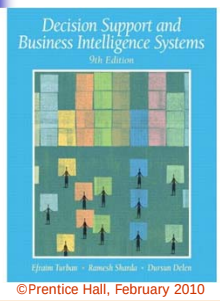
Dursun Delen, Ph.D.

Management Science & Information Systems
William S. Spears School of Business
Oklahoma State University

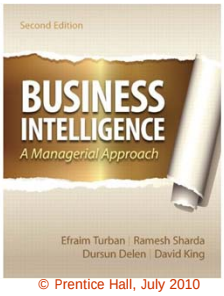
Presentation outline

- What is business intelligence?
- Why do we need it? Why now?
- What is knowledge discovery/data mining?
- What is the process of knowledge discovery?
- How does text mining differ from data mining?
- What are the best application areas for data and text mining (examples)?
- Where can I find more resources?

Primary sources...



©Prentice Hall, February 2010

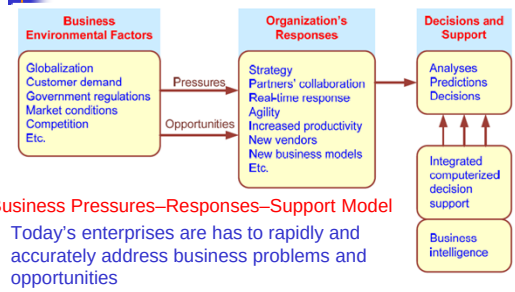


© Prentice Hall, July 2010

Business Intelligence (BI)

- BI is about getting the right information, to the right people, at the right time, in the right format for better decision making.
- BI is an umbrella term that combines tools, architectures, databases, analytics, etc.
- BI helps transform data, to information (and knowledge), to better decisions and actions.
- BI leads to:
  - fact-based decision making
  - "single version of the truth"

Why BI? Why now?

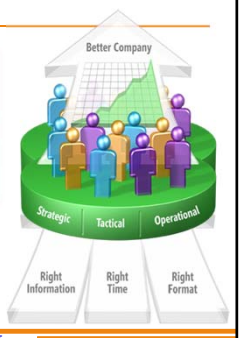


Business Pressures–Responses–Support Model

- ✓ Today's enterprises are has to rapidly and accurately address business problems and opportunities

Why BI? Why now?

Improving organizations by providing business insights to all employees leading to better, faster, more relevant decisions



- Advanced Analytics
- Self Service Reporting
- End-User Analysis
- Business Performance Management
- Operational Applications

A Brief History of BI

- The term BI was coined by the Gartner Group in the mid-1990s; however, the concept is much older
  - 1970s - MIS reporting - static/periodic reports
  - 1980s - Executive Information Systems (EIS)
  - 1990s - OLAP, dynamic, multidimensional, ad-hoc reporting -> coining of the term "BI"
  - 2005+ Inclusion of AI and Data/Text Mining capabilities; Web-based Portals/Dashboards
  - 2010s - yet to be seen

BI is an Umbrella Term

- ✓ BI includes reporting and analytics
  - ✓ Static versus dynamic/multi-dimensional reporting (e.g., OLAP/MOLAP)
  - ✓ Advanced analytics (e.g., data & text mining)

Why Data Mining? Why Now?

Needed

- More intense competition at the global scale.
- Needing to make accurate/timely decisions.
- Recognition of the value in data sources.

Available

- Availability of quality data on customers, vendors, transactions, Web, etc.
- Consolidation and integration of data repositories into data warehouses.
- The exponential increase in data processing and storage capabilities; and decrease in cost.

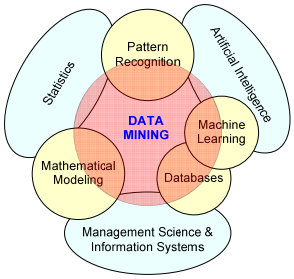
What is Data Mining?

- Data mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data stored in structured databases.

  - Fayyad et al., (1996)

- Data mining: a misnomer?
- Other names: knowledge extraction, pattern analysis, knowledge discovery, information harvesting, pattern searching, data dredging,...

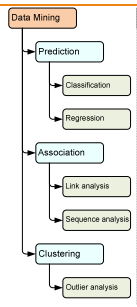
Data Mining at the Intersection of Many Disciplines



What Does DM Do?

- DM extract patterns from data
  - Pattern? A mathematical (numeric and/or symbolic) relationship among data items
- Types of patterns
  - Association
  - Prediction
  - Cluster (segmentation)
  - Sequential (or time series) relationships

A Taxonomy for Data Mining Tasks

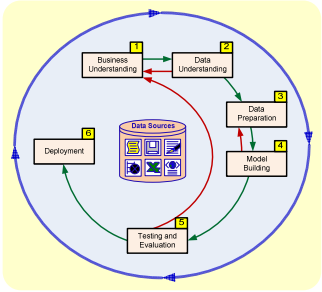


Learning Method	Popular Algorithms
Supervised	Classification and Regression Trees, ANN, SVM, Genetic Algorithms
Supervised	Decision trees, ANN/MLP, SVM, Rough sets, Genetic Algorithms
Supervised	Linear/Nonlinear Regression, Regression trees, ANN/MLP, SVM
Unsupervised	Apriory, OneR, ZeroR, Eclat
Unsupervised	Expectation Maximization, Apriory Algorithm, Graph-based Matching
Unsupervised	Apriory Algorithm, FP-Growth technique
Unsupervised	K-means, ANN/SOM
Unsupervised	K-means, Expectation Maximization(EM)

Data Mining Process

- A manifestation of best practices
- A systematic way to conduct DM projects
- Different groups has different versions
- Most common standard processes:
  - CRISP-DM (Cross-Industry Standard Process for Data Mining)
  - SEMMA (Sample, Explore, Modify, Model, and Assess)
  - KDD (Knowledge Discovery in Databases)

Data Mining Process: CRISP-DM



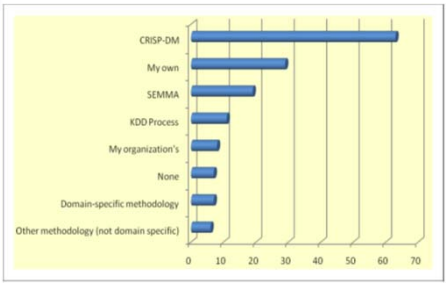
Data Mining Process: CRISP-DM

Step 1: Business Understanding
Step 2: Data Understanding
Step 3: Data Preparation (!)
Step 4: Model Building
Step 5: Testing and Evaluation
Step 6: Deployment

Accounts for ~85% of total project time

- The process is highly repetitive and experimental (DM: art versus science?)

Data Mining Process

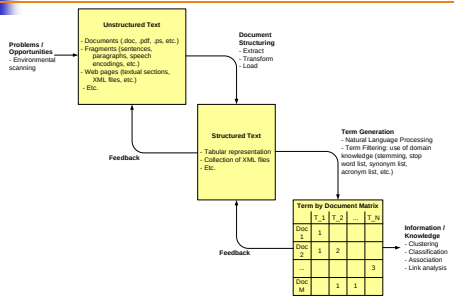


Source: KDNuggets.com, August 2007

Text Mining

- Text mining is a process that employs
  - a set of algorithms for converting unstructured text into structured data objects, and (Statistical Natural Language Processing)
  - the quantitative methods that analyze these data objects to discover knowledge (Data Mining)
- Text Mining = Statistical NLP + DM

Text Mining Process Map



Data Mining Application Areas

- Banking, finance/trading, insurance, retail, ...
- Science and engineering
- Government and defense
- Homeland security and law enforcement
- Healthcare
- Medicine

Highly popular application areas for data mining

- Entertainment industry
- Sports
- Etc.

Data Mining Applications... In Medicine and Healthcare

The main objective is to better understand the medical and healthcare problems/opportunities and provide data/fact driven knowledge for better and more timely decision making.

The idea is not to replace the traditional medical research but to augment it by providing new directions for clinical and biological research.

The resources

- The data is usually higher quality
- The reference literature is more organized
- Computational limitations are diminishing

Data Mining Applications... Military Health System (1/3)

- KBSI's Phase II SBIR research project
  - Funded through SBIR program by the Offices of the Secretary of Defense
  - SBIR: Phase I ⇒ Phase II ⇒ Phase III
- DM in Healthcare
  - Managerial
  - Clinical

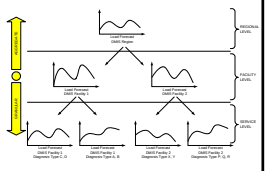
Reference: Delen, D. and S. Ramachandran (2003). A Hybrid Approach to Knowledge Discovery from Military Health Systems. Journal of Neural, Parallel and Scientific Computing, Volume 11, Number 1&2, pp. 161-183.

Data Mining Applications... Military Health System (2/3)

- One of the largest health systems in the US
  - 90+ hospitals
  - 100s of outpatient clinics, treatment facilities
  - 180,000 employees (doctors, nurses, other staff)
  - 8 million beneficiaries
  - $20 Billion/year budget
- Purpose/mission is to
  - Provide healthcare to eligible veterans
  - Provide education and training opportunities to health profession (residency, practical training, etc.)
  - Conduct medical research, create innovation
  - Provide public health service at the time of natural disasters
- Problem: ⇑ demand, ⇓ budget

Data Mining Applications... Military Health System (3/3)

- Classification/prediction
  - Demand forecasting
  - Resource allocation



- Association rules
  - Patient diagnosis

Data Mining in Medicine

- Analyzing large cancer databases...
  - Predicting and analyzing survivability for breast cancer [1] and prostate cancer [2] patients
- Methods/Materials:
  - Used artificial neural networks (MLP), decision trees (C5), logistic regression, support vector machines, ...
  - Used *k*-fold cross validation and more than 200,000 cases
- Results:
  - Predictions with better than 90% accuracy
  - Interesting and intriguing prognostic patterns

[1] Delen, D., G. Walker and A. Kadam (2004). "Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods", in ***Artificial Intelligence in Medicine***.
[2] Delen, D. and N. Patil (2006). "Knowledge Extraction from Prostate Cancer Data" in the proceedings of the ***39th Annual Hawaii International Conference on System Sciences*** (HICSS).
[3] Delen, D. 2009, "Analysis of Cancer Data: A Data Mining Approach," ***Expert Systems***, 26 (1), 100-112.

Data Mining in Medicine

- Analyzing organ transplantation data...
  - Discovering novel patterns from data to improve the organ transplantation process, save more lives...
- Methods/Materials:
  - Used variety of data mining techniques, including artificial neural networks, decision trees, support vector machines, logistic regression, and information fusion models ...
  - Used *k*-fold cross validation and more than 200,000 cases
- Results:
  - Highly accurate predictions on graft survivability
  - Interesting and intriguing prognostic patterns

[1] Delen, D., Oztekin, A., and Kong, Z. (2010) "A Machine Learning-Based Approach to Prognostic Analysis of Thoracic Transplantations," ***Artificial Intelligence in Medicine***.
[2] Delen, D., Oztekin, A., and Kong, Z. (2009) "Predicting the Graft Survival for Heart–Lung Transplantation Patients: An Integrated Data Mining Methodology," ***International Journal of Medical Informatics***.

A Brief List of Text Mining Applications

- Exploring trends in research literature
  - We did this for IS literature
- Identifying functional relationships between different genes using large collections of interdisciplinary research literature
- Analyzing records for customer contact centers
  - Both audio as well as text (SAS does this)
- Using text mining to detect deception
  - One of our Ph.D. student worked on this
- "Reading", summarizing classifying emails
- Anti-terrorism initiatives (CIA is using text mining)
- ...

Analysis of Research Literature with Text Mining - Data Collection

- Text mining of IS journals
  - ISR (Information Systems Research)
  - MISQ (MIS Quarterly)
  - JMIS (Journal of MIS)

Journal Name	Number of Articles	Dates of Publication	Volumes/Numbers Included
MIS Quarterly (MISQ)	246	1994 – 2005	18/1 – 29/3
Information Systems Research (ISR)	253	1994 – 2005	5/1 – 16/3
Journal of MIS (JMIS)	402	1994 – 2005	10/4 – 22/1
Total	**901**	**12 yrs**	

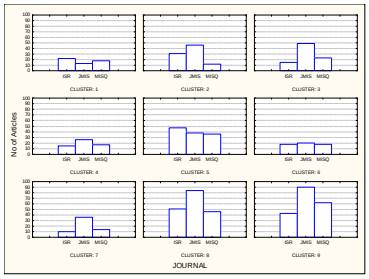
Analysis of Research Literature with Text Mining - Data Representation

Journal	Year	Author(s)	Title	Vol/No	Pages	Keywords	Abstract
MISQ	2005	A. Malhotra, S. Gosain and O. A. El Sawy	Absorptive capacity configurations in supply chains: Gearing for partner-enabled market knowledge creation	29/1	145-187	knowledge management supply chain absorptive capacity interorganizational information systems configuration approaches	The need for continual value innovation is driving supply chains to evolve from a pure transactional focus to leveraging interorganizational partner ships for sharing
ISR	1999	D. Robey and M. C. Boudreau	Accounting for the contradictory organizational consequences of information technology: Theoretical directions and methodological implications	2-Oct	167-185	organizational transformation impacts of technology organization theory research methodology intraorganizational power electronic communication mis implementation culture systems	Although much contemporary thought considers advanced information technologies as either determinants or enablers of radical organizational change, empirical studies have revealed inconsistent findings to support the deterministic logic implicit in such arguments. This paper reviews the contradictory
JMIS	2001	R. Aron and E. K. Clemons	Achieving the optimal balance between investment in quality and investment in self-promotion for information products	18/2	65-88	information products internet advertising product positioning signaling signaling games	When producers of goods (or services) are confronted by a situation in which their offerings no longer perfectly match consumer preferences, they must determine the extent to which the advertised features of
...	...	...	...	...	...	...	...

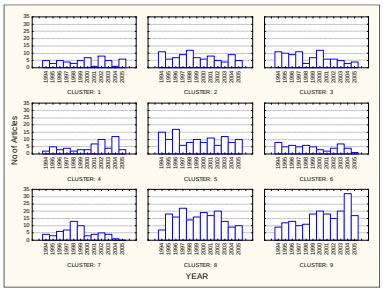
Analysis of Research Literature with Text Mining - Clustering Results

Cluster Number	Descriptive Terms (Top 10 Key words)	Article Frequency (Count)	(Percent)
1	Error, Discipline, MIS, Major, Methodology, Field, Value, Time, Future, Set	54	0.06
2	Network, Policy, Cost, Market, Team, Industry, Resource, Manager, Product, Structure	89	0.10
3	Executive, Financial, Competitive, Industry, Advantage, Investment, Market, Management, Business, Performance	87	0.10
4	Consumer, Price, Product, Customer, Online, Service, Site, Quality, Market, Cost	58	0.06
5	User, Database, Instrument, Model, Validation, Field, Construct, Development, Computer, Individual	121	0.13
6	Strategic, Innovation, Nature, Survey, Success, Management, Investment, Practice, Business, Empirical	56	0.06
7	Medium, Computer-Mediated, GSS, Communication, Task, Face-To-Face, Laboratory, Interaction, Idea, Social	60	0.07
8	Process, Change, Organizational, Practice, Case, Management, Method, Tool, Framework, Base	181	0.20
9	Risk, Project, Construct, Software, Investment, Factor, Manager, Behavior, Development, Value	195	0.22
	Total	**901**	**100**

Analysis of Research Literature with Text Mining - Clusters vs Journals



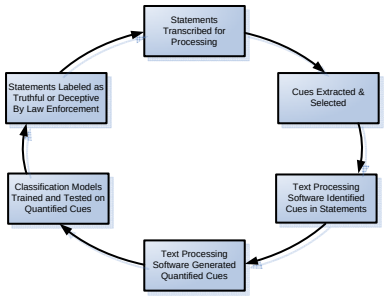
Analysis of Research Literature with Text Mining - Temporal Profiles



Deception Detection

- Deception: a message knowingly transmitted by a sender to foster a false belief or conclusion
- Commonly, deception research focused on cues
  - Nonverbal,
  - Paraverbal, and
  - Verbal (content cues)
- This research has to do with Automated Text-Based Deception Detection
  - Question: Using linguistic-based cues, can we distinguish truthful from deceptive messages in a high-stakes environment

Methodology



Methodology

This is a sample statement. Actual statements contain a suspect or witness description of a crime that he or she was involved in. These subject of the statement may be any type of crime ranging from very minor to a major event.
Each statement is transcribed as exactly as possible, including speling mistakes, grammar mistakes, and any punctuation used. There many be numerous mistakes in the actual statement.

GATE (text-mining tool)

Average Sentence Length	13.6
Average Word Length	4.941176470588235
Content Word Diversity	0.8157894736842105
Emotiveness	0.2903226806451613
Group References	0.0
Imagery	1.3929824561403512
Lexical Diversity	0.75

Overall Accuracy

Method	Cue Set 1 (Feature Selection - Linear Methods)	Cue Set 2 (Feature Selection - Literature)	Cue Set 3 (Feature Selection – ANN/Sensitivity)
ANN / MLP	74.85	73.62	73.01
ANN / RBF	75.46	69.94	74.23
Random Forest	69.33	75.46	76.07
Naïve Bayes	74.23	73.01	74.85
SVM	74.85	73.62	75.85

Forecasting Box Office Success of Hollywood Movies

Dursun Delen and Ramesh Sharda

Institute for Research in Information Systems
Oklahoma State University

Forecasting Box-Office Receipts: A Tough Problem!

"... No one can tell you how a movie is going to do in the marketplace... not until the film opens in darkened theatre and sparks fly up between the screen and the audience"

Mr. Jack Valenti
Long time President and CEO
of the Motion Picture Association of America

Current work on prediction...

- A lot of research have been done
  - Behavioral models
  - Analytical models
    - Predict after the initial release
- Our approach
  - Use a data mining approach
  - Use as much historical data as possible
  - Make it web-enabled
    - Predict before the initial release

Our Approach – Movie Forecast Guru

- DATA – 849 Movies released between 1998-2006
- Movie Decision Parameters:
  - Intensity of competition rating
  - MPAA Rating
  - Star power
  - Genre
  - Technical Effects
  - Sequel ?
  - Estimated screens at opening
  - ...
- Output: Box office gross receipts (flop → blockbuster)

Class No.	1	2	3	4	5	6	7	8	9
Range (in Millions)	< 1 (Flop)	> 1 < 10	> 10 < 20	> 20 < 40	> 40 < 65	> 65 < 100	> 100 < 150	> 150 < 200	> 200 (Blockbuster)

Data Description

Class No.	1	2	3	4	5	6	7	8	9
Range (in $Millions)	< 1 (Flop)	> 1 < 10	> 10 < 20	> 20 < 40	> 40 < 65	> 65 < 100	> 100 < 150	> 150 < 200	> 200 (Blockbuster)

Dependent Variable

Independent Variables

A Typical Classification Problem

Independent Variable	Number of Values	Possible Values
MPAA Rating	5	G, PG, PG-13, R, NR
Competition	3	High, Medium, Low
Star value	3	High, Medium, Low
Genre	10	Sci-Fi, Historic Epic Drama, Modern Drama, Politically Related, Thriller, Horror, Comedy, Cartoon, Action, Documentary
Special effects	3	High, Medium, Low
Sequel	1	Yes, No
Number of screens	1	Positive integer

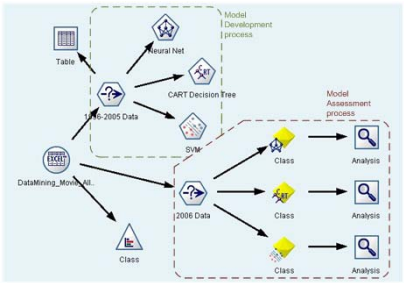
Our Approach – Movie Forecast Guru

PREDICTION MODELS

- Statistical Models:
  - Discriminant Analysis
  - Ordinal Multiple Logistic Regression
- Machine Learning Models:
  - Artificial Neural Networks
  - Decision Tree Induction
    - CART - Classification & Regression Trees
    - C5 - Decision Tree
  - Support Vector Machines
  - Rough Sets
  - ...

Data Mining Process Map

The DM Process Map in PASW



Prediction Results

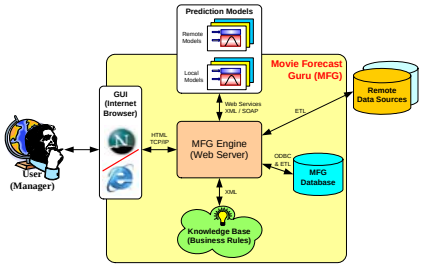
	Prediction Models					
	Individual Models			Ensemble Models		
Performance Measure	SVM	ANN	C&RT	Random Forest	Boosted Tree	Fusion (Average)
Count (*Bingo*)	192	182	140	189	187	**194**
Count (*1-Away*)	104	120	126	121	104	**120**
Accuracy (*% Bingo*)	55.49%	52.60%	40.46%	54.62%	54.05%	**56.07%**
Accuracy (*% 1-Away*)	85.55%	87.28%	76.88%	89.60%	84.10%	**90.75%**
Standard deviation	0.93	0.87	1.05	0.76	0.84	**0.63**

** Training set: 1998 – 2005 movies; Test set: 2006 movies*

Sample Prediction ...

Movie	Actual	No Category	War	Organization	Life	Character
Spider-Man 3	9	9	9	9	9	9
300	9	5	7	5	5	5
The Simpsons Movie	8	9	9	9	9	9
The Bourne Ultimatum	8	6	7	9	9	7
Knocked Up	7	5	5	5	5	7
Live Free or Die Hard	7	9	7	9	9	9
Evan Almighty	6	6	7	7	7	7
1408	6	5	5	4	5	5
TMNT	5	4	3	4	5	5
Music and Lyrics	5	4	5	5	5	5
Freedom Writers	4	4	3	4	4	5
Reign Over Me	3	4	4	4	4	4
Black Snake Moan	2	3	3	3	3	3
Ta Ra Rum Pum	1	1	1	1	1	1

Deployment: A Web-based DSS



Implementation of MGF

References for Movie Prediction

Media Coverage of "Box Office Forecasting" Project

- ✓ http://iris.okstate.edu/Movie_Media.htm

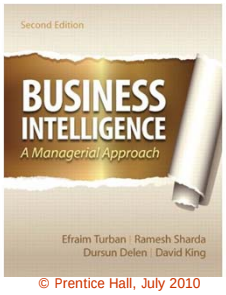
Delen, D., R. Sharda and P. Kumar (2006). "Movie Forecast Guru: A Web-based DSS for Hollywood Managers", Decision Support System. In Press.

Henry, M., R. Sharda and D. Delen (2007). "Using Neural Networks to Forecast Box-Office Success" America's Conference on Information Systems (AMCIS), Keystone, Colorado. Association for Information Systems, 1512-1516.

Sharda, R. and D. Delen (2006). "Predicting box-office success of motion pictures with neural networks" Expert Systems with Applications, 30(2), 243-254.

Sharda, R., D. Delen (2006). "How to Predict a Movie's Success at the Box Office", FORESIGHT: The International Journal of Applied Forecasting, October 2006.

The book that you need... Available.



Thank you for your attention

- Questions, comments, suggestions...

Contact Information
Dr. Dursun Delen
(918) 594-8283
dursun.delen@okstate.edu