Data Mining: Middleware or Middleman?

KDD Process Standards Panel

KDD-2000

August 20-23, 2000

Boston, MA, USA

Interactivity Track versus Invisibility Track

- KDD is defined as the “iterative” and “interactive” process of turning data into information and information into business solutions or intelligence
- “Successful technology becomes invisible” or “data mining (or KDD) will succeed when it becomes invisible”

Streamlining KDD Operations

- Standardized data models (or templates) that defines the most widely used data sources by industry - *vertical approach*
- Organize appropriate data mining operations/ tasks by field type - *horizontal approach*

Vertical Approach

A standardized data model for an e-commerce operation may include:

- ***Membership table*** (user-level) that includes name, shipping and billing information, profile and preference information, privacy indicators, promotion information, coupon redemption, and personal demographics acquired through on-site registration, etc.
- ***URL or click-stream table*** (page view level) that includes URL trails (visit trails), URL type (post-purchase, pre-purchase, search, etc.), search results, keywords, etc.
- ***Other views/ tables*** such as Affiliate table (user level), Domain table (affiliate level), Ad table (hit level), Message table (page view level), Promotion table (page view level), Site Content and Structure table, Administrative table (hit level,) etc.

Horizontal Approach

- Organize KDD tasks by type of data element found (such as nominal, ordinal, interval or continuous, etc.)
- Let subsequent KDD operations (such as data preprocessing, exploration, mining, etc.) roll based on this information
- Data mining algorithms can be pre-configured and optimized to solve a specific business problems

Next Steps

- Set-up a task force to study existing KDD process standardization efforts
- Sign-up sponsors from the industry (e.g., retail banking, telecommunications, e-commerce) willing to provide representative data sets
- Sign-up database and data mining vendor support
- Once formalized, introduce and make these efforts part of the data mining/ KDD curriculum

Panel Participants

- CPEX (Customer Profile Exchange) offers a vendor-neutral, XML-based open standard for facilitating the privacy-enabled interchange of customer information across disparate enterprise applications and systems. [http://www.cpex.org]
- CRISP-DM is an industry consortium developing an industry-neutral and tool-neutral CRoss-Industry Standard Process Model for Data Mining. [http://www.crisp-dm.org/]
- DMG, the Data Mining Group, is a consortium of industry and academics formed to create standards, starting with PMML, (XML-based) for defining and sharing predictive models. [http://www.dmg.org]

Panel Participants

- MDC (Meta Data Coalition) regroups vendors and users allied with a common purpose of driving forward the definition, implementation and ongoing evolution of a meta data interchange format and its support mechanisms. [http://www.MDCinfo.com/]
- OLEDBDM (OLE DB for Data Mining), a Microsoft effort extending SQL databases through a new API to better support data mining operations. [http://research.microsoft.com/dmx/oledbdm/]
- TASF (The Analytical Solutions Forum), an industry consortium whose mission is to establish solution-oriented performance criteria and interoperability requirements within and between classes of decision support tools (including but not limited to OLAP or On-line Analytical Processing, data mining, data visualization, text processing, and decision analysis). [http://www.tasf.org/]