KDD Process Standards Panel

Position Statement

Ismail Parsa, Panel Chair, KDD Process Standards

Data Mining: Middleware or Middleman?

On the one hand, KDD is defined as the "iterative" and "interactive" process of turning data into information and information into business solutions or intelligence.[1] On the other hand, this notion that "successful technology becomes invisible" or "data mining (or KDD) will succeed when it becomes invisible"[2] has been coming from the KDD community for a while. I call these tracks the *interactivity track* and the *invisibility track*, respectively. But by definition, these two tracks are at odds with one another: KDD can not be interactive[3] and invisible at the same time. A few will doubt the validity of the argument behind the invisibility track: data mining will find mainstream acceptance when it gets embedded in our lives just like a utility or a household appliance. Today, however, the process of knowledge discovery and data mining is far from automated (i.e., still requires human interaction) and, therefore, is difficult to deploy effectively. Many commercial tools lack functionality and scalability that customers require.

A sound way around this issue -- well known to systems integrators/ Customer Relationship Management (CRM) consultants -- is to streamline KDD operations through the use of standardized data models. These data models -- typically organized by vertical industry -- include the most widely used data sources in an industry. For example, a standardized data model for an e-commerce operation may include the following tables/ views:

- 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.

---

[1] For a more formal definition, please see pp. 4-11, Fayyad, Usama M., Gregory Piatetsky-Shapiro, Padhraic Smyth and Ramasamy Uthurusamy, 1996. *Advances in Knowledge Discovery and Data Mining*. M.I.T. Press.

[2] For more on this, please see Usama Fayyad's editorial article and/ or George John's article in the first issue of KDD Explorations (http://research.microsoft.com/datamine/..)

[3] I also take the view that "iterative" processing is, to a great extend, a by-product of interactivity.

- 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.

Similarly, a standardized data model for financial services may include customer demographics, promotion and response information, channel information, credit information, card usage information, etc.

The key to streamlining the KDD tasks is to clearly define in the metadata the type for each data element found in these views/ tables (such as nominal, ordinal, interval or continuous, etc.) and to let subsequent KDD operations (such as data preprocessing, exploration, mining, etc.) roll based on this information. In this setting, the KDD process (including the data mining algorithms) can be pre-configured and optimized to solve a specific business problem, such as predicting churn in the telecommunications industry or estimating the size of the orders in retail/ catalogue industry.

Once formalized, these data models can be published as industry-specific metadata standards (XML-based?) and can be deployed by database and/ or data mining solution vendors. I see it upon this community to own the invisibility track instead of the systems integrators or the CRM consultants. Because if we don't, we continue being the middleman we are and not the middleware we should become. Toward this goal, the SIGKDD organization can embrace the following:

- Set-up a task force to study existing KDD process standardization efforts and to settle on one or more of these solutions
- Sign-up sponsors from the industry (e.g., retail banking, telecommunications, e-commerce) willing to provide representative data sets to turn concepts into practical solutions
- Sign-up database and data mining vendor support
- Introduce and make these efforts part of the data mining/ KDD curriculum in schools – perhaps, this topic can be more formally covered in the human interaction and background knowledge steps of the KDD process

A standardized process for data mining clearly offers the window of opportunity for data mining to "go behind the scenes" and to become invisible in our lives, much as a utility or a household appliance. Not having it is much like establishing a country, a political system without a constitution. This panel brings together the

representatives of institutions/ on-going efforts working independently toward streamlining whole or part of the KDD and/or related processes. Our objective is to inform the KDD community on the current state of their work and, possibly, to open doors for mutual collaboration. Among them are (in alphabetical order of occurrence):

- 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]

- 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/]