# Feeding the Knowledge Base: Data Analysis, Data Mining and Case-Based Retrieval Systems

Freda Salatino

## Congratulations!

You've managed to convince The Powers That Be that the best way to leverage your corporation's legacy of technical data, is by creation of a knowledge system.

Welcome to The Bleeding Edge™. Esther Dyson would be proud of you. Give yourself a round of applause.

(5 seconds, please)

NOW what do you do?

## What's Needed?

The knowledge system will contain legacy content, content that originates in-house, and content that originates through interaction with customers.

But knowledge is not a self-renewing commodity.

To keep your knowledge system useful, you'll have to create a means to constantly tune, refine, and augment its information, so that it can keep up with -- and maybe even ahead of -- the needs of its clientele.

## Crafting The Knowledge Base

One of the best methodologies for creating new knowledge out of a morass of legacy information is called data mining.

Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data, using system-driven methods such as pattern recognition and statistical techniques.

## Crafting the Knowledge Base

Data mining technologies provide the following capabilities:

- Automated discovery of previously unknown patterns
- Automated prediction of trends and behaviors

## Data Mining vs. Data Analysis

Data analysis is best used when there is a prior hypothesis to be investigated. The analyst investigates only those parts of the data that they believe will produce the key relationship.

Data mining tools "drill down" into the data, exposing trends, relationships and patterns that may have previously been unknown.

## Data Mining vs. Data Analysis

Thus, data mining tools enable **knowledge discovery,** rather than knowledge verification.

## Data Mining Methodologies

- Clustering -- identifies groups of closely-related records according to statistical similarities. Algorithms automatically assign records with a large number of similar attributes into a relatively small set of groups. Can identify a previously unrealized set of customers/products/problems that have similar attributes.)
- Classification -- implies the defining characteristics of a dataset. Models created from a set of classified examples, are used predictively, to assigns new records to existing classes.

- Association -- uses rules to determine the relationship between events that occur at one time. Particularly useful in in market-based analysis.
- Sequence-based analysis (temporal patterns) -- used to identify time-based affinity; like association, but emphasizes cases where time is the dominating dimension in the data.
- Prediction/forecasting -- the creation of models using data mining algorithms, to predict the likelihood of a future event.

## Data Mining Techniques

- Rules-based analysis (rule induction) -- the reasoning from specific facts to reach an hypothesis. Enables the decision tree to “drill down” through existing data without the need for a prior hypothesis.

  *This comes into serious play in Case-based Reasoning systems.*
- Neural networks -- predictive models that “learn” how to solve problems based on examples; best used to model non-linear data, or data that is missing some values.

- Classification and regression trees -- decision tree technique used for classification of a dataset. Fosters rules that can be applied to a new (unclassified) dataset to predict which records will have a given outcome.
- Nearest neighbor analysis -- classifies each record in a dataset based on a combination of the records most similar to it in a historical dataset.

- Decision trees -- tree-shaped structures that represent a set of decisions. Best suited to classification and clustering tasks.
- Data visualization -- enables the user to detect and comprehend patterns and possible anomalies, but do not actually “automatically” discover previously unknown patterns and trends. Used mainly to help the user comprehend the results of data mining.

## Feeding the Knowledge Base

One of the best methodologies for “feeding” and interacting with an existing knowledge base is called case-based reasoning.

Case-based reasoning combines statistical analysis and deductive logic to create a structured “path” through an existing knowledge base. Users pose an initial question, adding qualifiers to hone their search until they reach a solution with a high statistical probability of answering their question.

## Feeding the Knowledge Base

Each question-answer pair is considered a separate “case.”

Each question that hits a “dead end” in your knowledge base points out a need for further information.

## Case-based Reasoning Systems

- Case-based Reasoning (CBR) systems are constructed to cover one or more domains, or possible areas of investigation.
- Each domain contains a forest of decision trees. Access to a particular decision tree is set by a series of rules.
- Each decision tree contains branches that terminate in one specific case.

## Lather, Rinse, Repeat

- If the search ends in an answer and a fix, you’ve prevented a support call, empowered the user, AND told the user that your product is easy to use.
- If the user finds an answer but NOT a fix, you find a hole in your knowledge system -- and/or your product testing.
- If the user finds neither an answer nor a fix, you still haven’t lost anything.

## Lather, Rinse, Repeat

In any case, the user is helping you sustain, correct, and grow the knowledge base.

## Exercise #1: Loaves Into Fishes

- Customer support sites all use the kind of deductive logic used to construct CBR domains and decision trees. The FAQs themselves are organized by assumptions (rules) the company support staff have made about their customers.
- Let’s return to the Adobe support site and see if we can turn last week’s Usability Test into an actual Case.
- What is the domain? What are the decision trees? What are the rules?

## Exercise #2: What's Missing?

- Since many CBR systems (in fact, most self-guided User Assistance) are created by mining legacy data, there are bound to be new cases that elude detection -- until some user complains that they didn't turn up an answer during their search.
- Let's see how the discovery of new cases, forces the developer to change the rules for a CBR system.

## Attributes of Case-based Data

- Most CBR tools let users make physical connections between strings of data, constructing a decision tree of data as objects.
- As with any other data retrieval system, these objects must be assigned attributes, or the rules of the CBR system will not be effective.
- What are some of the attributes of the information we just went through on the Adobe site?
- What other characteristics should "case fodder" have, in order to be effective in a CBR system?

## Self-guided User Assistance Systems To Visit

### Non-Automated

- http://help.netscape.com
- http://www.adobe.com/supportservice/custsupport/
- http://www.amazon.com/exec/obidos/subst/help/desk.html/002-7587996-8220000
- http://pages.ebay.com/aw/help/help-start.html

### Automated

- http://seudo.broder.com/intro/introt.htm
- http://www.lucasarts.com/support/