Mining in the Global Brain

Krzysztof (Krys) Cios

October 10, 2002
Discovering Knowledge on WWW

World Wide Web can be thought of as Global Brain

- Web has about $3.0 \times 10^9$ web pages

- Human brain has about $10^{10}$ neurons

- Web pages are highly interconnected
  - The knowledge is hidden in:
    - Individual web pages
    - The structure/interconnections

- It is impossible to analyze the Web without resorting to machine intelligence
What useful knowledge can be learned from these two independent pieces of information?
Machine Intelligence Tools

Data Mining - concerned with extraction of valid, useful, easily understandable knowledge from large collections of data, for high level decision making

Machine Learning – generation of a new data structure that is different from an old one
  • generates knowledge in the easy to understand format of IF…THEN… rules

Data Integration - methods that provide unified access to semantically and structurally different information sources

Cios, K.J., Pedrycz, W., Swiniarski, R., Data Mining Methods for Knowledge Discovery, Kluwer, 1998
What are we looking for?

The knowledge hidden in individual and/or different web pages
- Information Integration methods discover correspondence between data stored in different web pages
- It can be done using Machine Learning
Structured vs. Unstructured Data

**Structured XML data**
- documents containing structured information
- easy to define semantics of the stored information

**Unstructured/Semi-structured HTML or raw text data**
- information stored in unstructured way
- hard to identify meaning (semantics) of the information

---

**Rule Induction Papers**

R.S., Michalski, I., Mozetic, J., Hong, H., Lavrac, (1986)
"The multipurpose incremental learning system AQ15 and its testing application to three medical domains", Proceedings of the Fifth National Conference on Artificial Intelligence, Morgan-Kaufmann, Philadelphia, PA, 1041-1045

J.R., Quinlan, (1990)
"Learning logical definitions from relations", Machine Learning, 3, 239-266

"CLIP: cover learning using integer programming", Kybernetes, 26(4-5)

P., Clark, T., Niblett, (1989)
"The CN2 Algorithm", Machine Learning, 3, 261-263

---

**Structured vs. Unstructured Data**

Structured XML data
- documents containing structured information
- easy to define semantics of the stored information

Unstructured/Semi-structured HTML or raw text data
- information stored in unstructured way
- hard to identify meaning (semantics) of the information
How to discover the knowledge?

Data Integration methods are highly accurate and can be used for discovering large amounts of interconnected information on the WWW

- Structured documents can be easily converted into format compatible with most data integration methods
  - XMapper algorithm

Additional factors weakening the capacity for effective port control are the myriad of local vessels in the port and the often diverse geography of the surrounding area. Barges, tankers, launches, lighters, patrol craft and tugs are only a few of the vessels with more than one reason for being in the port and going alongside ships that are anchored, docked or moored through the port. It does not require much imagination to devise smuggling schemes involving service vessels. Yacht basins, fish factories, abandoned piers and warehouses provide traffickers with plenty of opportunities for unloading drugs from vessels. Yachts and fishing vessels can quickly approach a ship and procure a drug shipment from it. Alternatively, a drug consignment may be thrown overboard in a buoyant container and retrieved by a boat waiting nearby.

Another related topic concerns the illicit activity of personnel serving in an official capacity on the docks or having some other legitimate reason for being there. Ports and their surrounding areas have been the setting for some of the most blatant forms of illegal activity, ranging from single-handed pilferage to highly sophisticated, organized criminal enterprises. Among the diverse personnel exposed to such opportunities are cargo handlers, longshoremen, seamen, port police, security guards, shipping agents and truck drivers. The ready and legitimate access these individuals have to ships enables them to take drugs ashore with less risk of being noticed.
How to discover the knowledge?

Data Mining methods are fast and accurate

- Structured documents can be easily converted into format compatible with most DM methods
  - 4cRuleBuilder and DataSqueezer algorithms

IF
“yachts, fishing boats, small barges, tugboats, lighters, launches are OK” AND
“preferred professions: cargo handlers, longshoremen, seamen, port police, security guards, shipping agents and truck drivers” THEN
hired by Joe Doe

IF
“barges, tankers, launches, lighters, patrol craft and tugboats” AND
“cargo handlers, immigration and Customs officers, longshoremen, seamen, port police, security guards, shipping agents and truck drivers” THEN
drug trafficking
Scenario Uncovered

**IF**

hired by John Doe

THEN

possible drug trafficking

Only by integration of the dispersed web information and by using data mining methods on the integrated data we can hope for automating the discovery of useful knowledge for quick decision making.
Dilemma

ideas / entrepreneurship / economy

security