



University of Colorado at Denver
University of Colorado at Boulder
University of Colorado Health
Sciences Center
Center for Computational Biology



4cData, LLC
for See Your Data



Mining in the Global Brain

Krzysztof (Krys) Cios

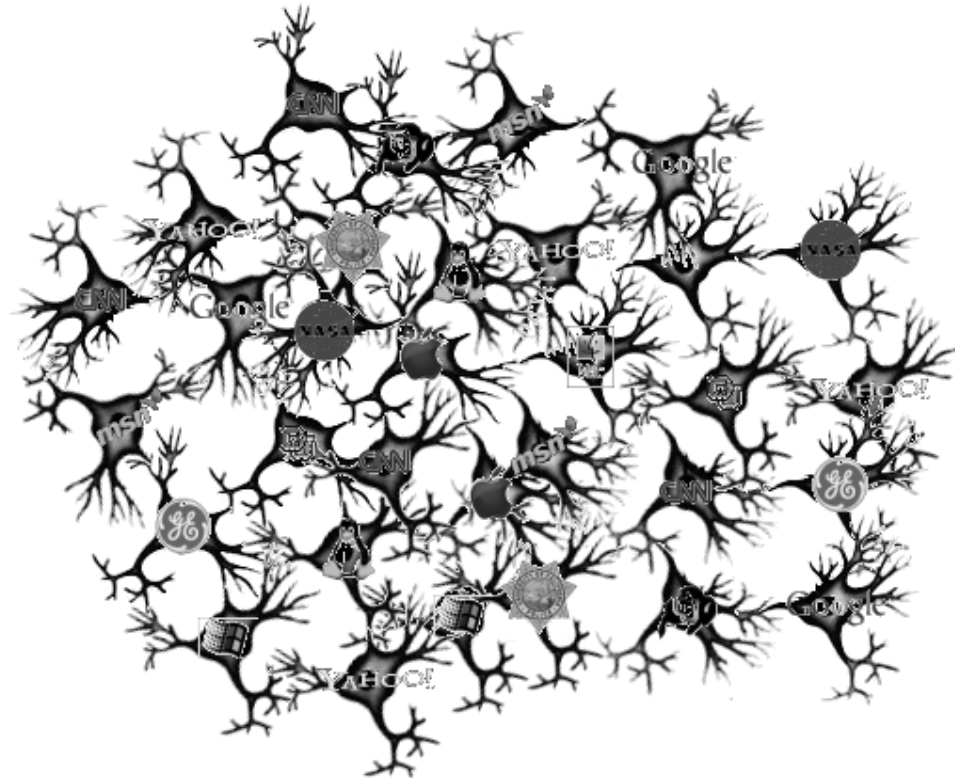


October 10, 2002

Discovering Knowledge on WWW

World Wide Web can be thought of as Global Brain

- Web has about 3.0×10^9 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



Cios

Scenario

online article

online job ad



UNODCCP United Nations Office for Drug Control and Crime Prevention

Home Site map Links Contact us Field Offices select the site

October 5, 2002

Ports and the areas surrounding them

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 tugboats are only a few of the vessels with bona fide reasons for being in the port and going alongside ships that are anchored, docked or moving 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, immigration and Customs officers, 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.

ALL SECTIONS

Employment

8.20.2002 ☺

Joe Doe

☎ N/A

🌐 great_money@free.email.com



we are looking for young man:

- wanting to earn great money
- must have flexible schedule
- must be a boat owner
 - yachts, fishing boats, small barges, tugboats, lighters, launches are OK
- preferred professions: cargo handlers, longshoremen, seamen, port police, security guards, shipping agents and truck drivers

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

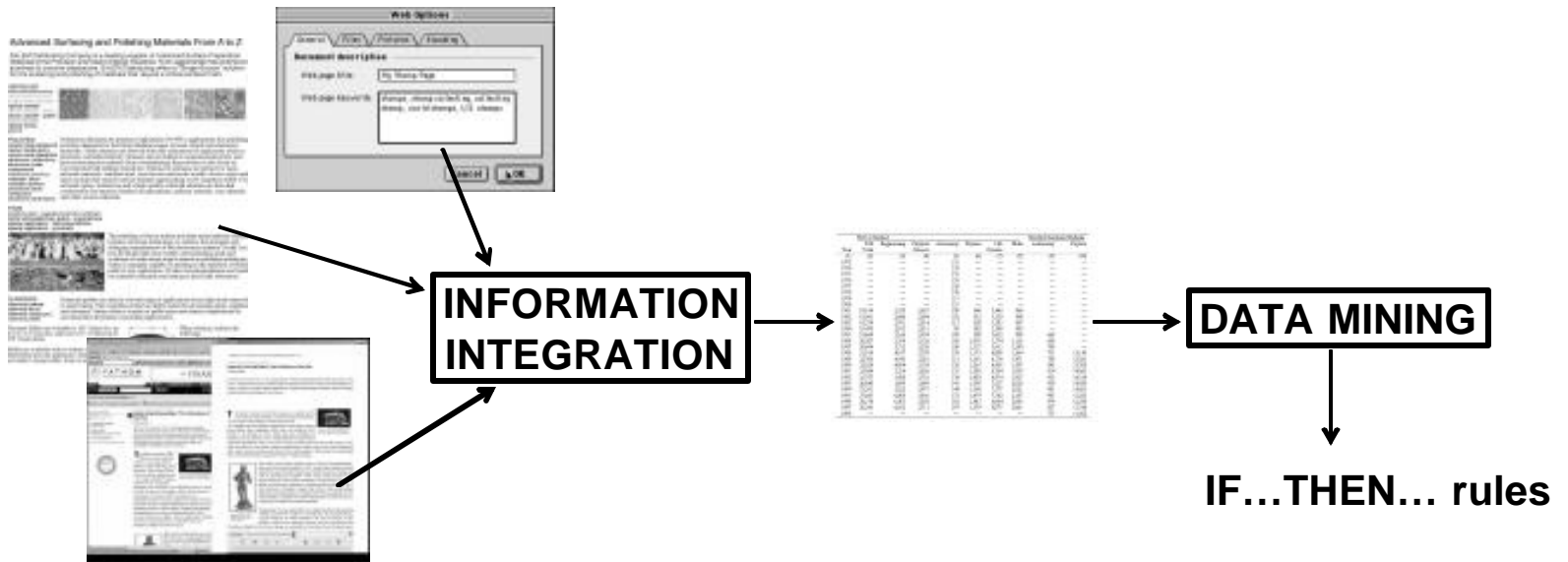
- generates knowledge in the easy to understand format of **IF...THEN... rules**

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

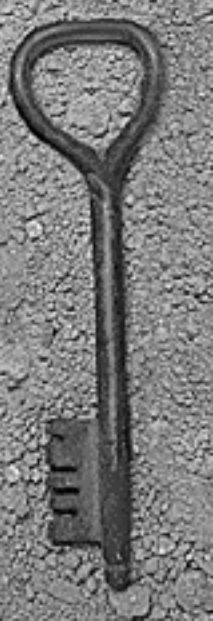
What are we looking for?

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



Rule Induction Papers

R.S., Michalski, I., Mozetic, J., Hong, N., 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*, 5, 239-266

K.J. Cios, D.K., Wedding, N., Liu, (1997)

"CLIP3: cover learning using integer programming", *Kybernetes*, 26(4-5)

P., Clark, T., Niblett, (1989)

"The CN2 Algorithm", *Machine Learning*, 3, 261-283

```
<?xml version="1.0" standalone="no" ?>
```

```
<MLBibliographies>
```

```
  <PageTitle>Rule Induction Papers</PageTitle>
```

```
  + <Publication N="1">
```

```
  - <Publication N="2">
```

```
    - <Author>
```

```
      <Firstname>J.R.</Firstname>
```

```
      <Lastname>Quinlan</Lastname>
```

```
    </Author>
```

```
    <Year>1990</Year>
```

```
    <Title>Learning logical definitions from relations</Title>
```

```
    <Journal>Machine Learning</Journal>
```

```
    <Volume>5</Volume>
```

```
    <Conference />
```

```
    <Publisher />
```

```
    <Location />
```

```
    <Pages>239-266</Pages>
```

```
  </Publication>
```

```
<html>
```

```
<head>
<meta http-equiv="Content-Type" content="text/html;
charset=windows-1252">
<meta name="GENERATOR" content="Microsoft FrontPage 4.0">
<meta name="ProgId" content="FrontPage.Editor.Document">
<title>Rule Induction Papers</title>
</head>
```

```
<body>
```

```
<p style="PADDING-RIGHT: 4px; PADDING-LEFT: 4px; FONT-WEIGHT:
bold; PADDING-BOTTOM: 4px; COLOR: white; PADDING-TOP: 4px;
BACKGROUND-COLOR: teal"><font face="Arial" size="5">Rule
Induction Papers</font></p>
<p style="BACKGROUND-COLOR: white"><font face="Arial"
color="red" size="1">Maintained
by Lukasz Kurgan<br>
<br>
</font></p>
<p style="MARGIN: 0px; BACKGROUND-COLOR: white"><font
face="Arial" color="black" size="3"><b>R.S.,
Michalski, I., Mozetic, J., Hong, N., Lavrac,
(1986)</b></font></p>
```

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



ALL SECTIONS

Employment
 8.20.2002 ☺ **Joe Doe** ☎ N/A 🌐 great_money@free.email.com ✉

we are looking for young man:

- wanting to earn great money
- must have flexible schedule
- must be a boat owner
 - yachts, fishing boats, small barges, tugboats, lighters, launches are OK
- preferred professions: cargo handlers, longshoremen, seamen, port police, security guards, shipping agents and truck drivers

Ports and the areas surrounding them

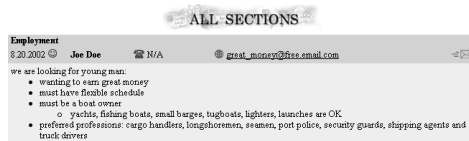
Additional factors weakening the capacity for effective port control are the myriad of local vessels in the port and the often diverse geography of the surrounding area. Barges, tankers, launches, lighters, patrol craft and tugboats are only a few of the vessels with bona fide reasons for being in the port and going alongside ships that are anchored, docked or moving 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, immigration and Customs officers, 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

ALL SECTIONS

Employment
8/30/2002 Joe Doe @ great_money@frees.email.com

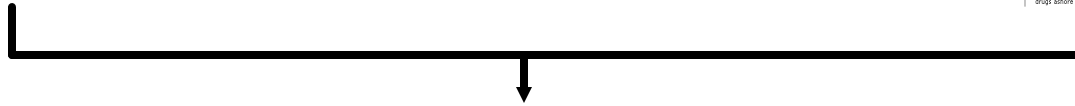
we are looking for young man

- wanting to earn great money
- must have flexible schedule
- must be a boat owner
 - yachts, fishing boats, small barges, tugboats, lighters, launches are OK
- preferred professions: cargo handlers, longshoremen, seamen, port police, security guards, shipping agents and truck drivers

UNODC
Port and the areas surrounding them

Additional factors weakening the capacity for effective port control are the myriad of local vessels in the port and the other diverse geography of the surrounding area. Barges, tankers, launches, lighters, patrol craft and tugboats are only a few of the vessels with bona fide reasons for being in the port and going alongside ships that are anchored, docked or moving through the port. It does not require much imagination to devise smuggling schemes involving service vessels. Tug boats, tug tenders, abandoned piers and warehouses provide traffickers with plenty of opportunities for unloading drugs from vessels. Tugboats and fishing vessels can quickly approach a ship and procure a drug shipment from it. Alternatively, a drug shipment may be thrown overboard in a buoyant container and retrieved by a boat waiting nearby.

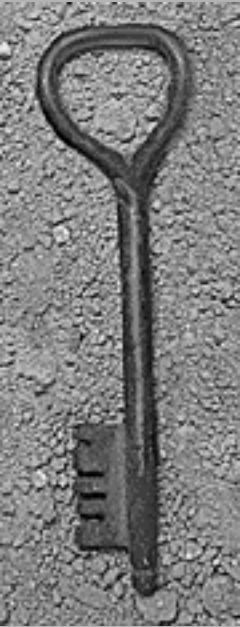
Another related topic concerns the illicit activity of personnel serving in an official capacity on the docks or having some other legitimate reason for being there. Ports and their surrounding areas have been the setting for some of the most blatant forms of illegal activity, ranging from single-handed efforts to high sophisticated, organized criminal enterprises. Among the diverse personnel exposed to such opportunities are cargo handlers, immigration and Customs officers, 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.



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.

Cios



Dilemma

