Strengthening Collaborative Data Analysis and Decision Making in Web Communities

Nikos Karacapilidis*, Spyros Christodoulou*, Manolis Tzagarakis*, Georgia Tsiliki** and Costas Pappis***

* CTI & University of Patras, Greece
** Biomedical Research Foundation, Academy of Athens, Greece
*** University of Piraeus, Greece

WI&C’14 Workshop @ WWW 2014
April 7, 2014 - Seoul, S. Korea
the Dicode concept

http://www.flickr.com/photos/43343237@N04/5593973222/
Basic issues in Dicode

data
intensiveness
cognitive
complexity

big
data

sense
making

variety
volume
velocity

data

human
reasoning

machine
reasoning

decision
making
“In spite of the tremendous advances made in computational analysis, there remain many patterns that humans can easily detect but computer algorithms have a hard time finding”

“Ideally, analytics for Big Data will not be all computational – rather it will be designed explicitly to have a human in the loop”

“With Big Data, the use of separate systems becomes prohibitively expensive ... Big Data has made it necessary to run heterogeneous workloads on a single infrastructure that is sufficiently flexible to handle all these workloads”.

“It is rarely enough to provide just the results. Rather, one must provide supplementary information that explains how each result was derived, and based upon precisely what inputs. Such supplementary information is called the provenance of the (result) data”.

“Systems with a rich palette of visualizations become important in conveying to the users the results of the queries in a way that is best understood in the particular domain”.

Divyakant Agrawal, UC Santa Barbara
Philip Bernstein, Microsoft
Elisa Bertino, Purdue Univ.
Susan Davidson, Univ. of Pennsylvania
Umeshwar Dayal, HP
Michael Franklin, UC Berkeley
Johannes Gehrke, Cornell Univ.
Laura Haas, IBM
Alon Halevy, Google
Jiawei Han, UIUC
H. V. Jagadish, Univ. of Michigan (Coordinator)
Alexandros Labrinidis, Univ. of Pittsburgh
Sam Madden, MIT
Yannis Papakonstantinou, UC San Diego
Jignesh M. Patel, Univ. of Wisconsin
Raghu Ramakrishnan, Yahoo!
Kenneth Ross, Columbia Univ.
Cyrus Shahabi, Univ. of Southern California
Dan Suciu, Univ. of Washington
Shiv Vaidyanathan, IBM
Jennifer Widom, Stanford Univ.
## Big Data challenges

<table>
<thead>
<tr>
<th>Challenge</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lack of skills/expertise needed to run analysis on all the data</td>
<td>57%</td>
</tr>
<tr>
<td>Too difficult to access all data and make available to users for analysis</td>
<td>50%</td>
</tr>
<tr>
<td>Not effectively using our most valuable data to drive decisions</td>
<td>45%</td>
</tr>
<tr>
<td>Too difficult to analyze and understand all of the data</td>
<td>37%</td>
</tr>
<tr>
<td>Too difficult to share information and insights with others</td>
<td>22%</td>
</tr>
<tr>
<td>Running queries and reports takes too long</td>
<td>19%</td>
</tr>
</tbody>
</table>

Source: IDG Research Services, August 2012
Main goal

• What
  – facilitate and augment collaboration and decision making in data-intensive and cognitively-complex settings

• How
  – by exploiting and building on the synergy of human and machine reasoning
  – by deepening our insights on the proper exploitation of big data and related technologies
Synergy of machine & human reasoning
Achievements
The Dicode architecture
Collaboration workspaces

- Collection & sharing of collaboration items
- Interrelation & evolution of collaboration items
- Informal / semiformal argumentation
- Informal / semiformal aggregation of collaboration items
- Formal exploitation of collaboration items patterns
- Formal argumentation and reasoning mechanisms

Formality
The Dicode Workbench
Dicode Services

• Collaboration and Decision Making Support services

• Forum Summarization service
  – takes a cluster of discussion threads (from various web forum types) as input and identifies the most prominent terms (topics)

• Subgroup Discovery service
  – searches for subgroups in any user provided data by searching the rules that cover many target value examples and few non-target value examples

• Blog-preprocessing service
  – returns a condensed representation of weblog entries containing only significant nouns

• Topics service
  – aims to give the user a quick overview of the thematic content of a document collection

• Keytrends service
  – returns metadata about Tweets on a selected day; in each case, the top 100 (or less) values are returned.

• Twitter preprocessing service
  – returns a condensed representation of Tweets containing only significant nouns. Optionally, one can query for a search term (a regular expression).
Collaboration Service: ‘mind-map view’
Collaboration Service: Calling a data mining service
Synergy of human and machine reasoning
Collaboration workspace: Filtering items
Collaboration workspace: Proactive help
Collaboration service: ‘formal view’

- Comparative NGS vs Gene Expression Tamoxifen data
  - Work with gene-expression data
  - Augment the genomic data
- Consider also NGS data
  - Unfamiliar with the technology
  - Provide representative dataset
  - Confront ambiguities
- Notable agreement between NGS and GE analyses
  - SD for Tam data
  - SD for Tam RNA-Seq
  - Important publications

[Screen capture of a browser window showing a collaboration service interface with tasks and comments]
Workspace analytics view
concluding ...
Advancing the state-of-the-art

- **Synergy of human and machine reasoning**
  - Exploitation of collective intelligence in parameterizing data mining solutions, selecting the appropriate data sources, and interpreting data mining outputs
  - Preserving provenance of decision making
  - Appropriate visualizations of collaboration towards problem solving

- **Practical success of data mining solutions**
  - enabling users to guide and control the data mining process and include their domain knowledge
  - more compact and semantically-enriched data mining results
  - the overall usability of the data mining system, in particular the ability to re-use existing solutions and built upon documented decision provenance
Big Data challenges (revisited)

User-friendly environment; No data-mining expertise

Easy access through the Dicode Workbench and integrated services

Semantically-enriched data; Knowledge-based decision making

Multiple collaboration views; Argumentation-based reasoning

Intelligent data mining & collaboration support solutions; Integrated reporting functionalities

Dicode’s Collaboration Services; Exploiting the competences of all stakeholders to meaningfully confront various information management issues
The Dicode project - http://dicode-project.eu/

DICODE – “Mastering Data-Intensive Collaboration and Decision Making” is co-funded by the European Commission under the contract FP7-ICT-257184.
Mastering Data-Intensive Collaboration and Decision Making

Research and Practical Applications in the Dicode Project

Nikos Karacapilidis - Editor
Thanks for your attention!

Nikos Karacapilidis
CTI & University of Patras, Greece
nikos@mech.upatras.gr